The Complex Third-Party Tracking Ecosystem: A Multi-Dimensional Perspective

Marjan Falahrastegar

A thesis submitted to the University of London in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

School of Electronic Engineering and Computer Science

Queen Mary University of London

United Kingdom

October 2016

Dedicated to Maman, Baba, Shervin and Hossein

Statement of Originality

I, Marjan Falahrastegar, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged below and my contribution indicated. Previously published material is also acknowledged below.

I attest that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge break any UK law, infringe any third party's copyright or other Intellectual Property Right, or contain any confidential material.

I accept that the College has the right to use plagiarism detection software to check the electronic version of the thesis.

I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university.

Signature:

Marjan Falahrastegar

Date:

Acknowledgements

I would like to thank my dear supervisor Steve Uhlig for his positiveness, generous help, continuous encouragement, and his vast perspective on the complex human ecosystem which its complication is far beyond the ecosystem studied in this thesis. Additionally, I appreciate any constructive feedback that I have received from Richard Mortier and Hamed Haddadi.

I would like to thank all my friends and colleagues at Queen Mary University of London, Sabri, Amna, Shan, Jie, Kishan, Timm, Ignacio, Eder, Felix and Gareth. We have shared so much laughter and sometimes tears during these years. Additionally, I would like to thank all other friends who are scattered around the world, in particular Mahsa.

I cannot thank enough my parents, Mahvash and Mohammad Reza, and my dear brother, Shervin, who without their unconditional love and support this journey would have never begun. They have sacrificed a lot for me to be here and I will remain indebted to them forever. Last but not least, my dear husband Hossein, who has been my secret supervisor, first-line reviewer and all-the-time companion through this journey. I am blessed to have him in my life.

Abstract

The third-party tracking ecosystem continuously evolves in scope, therefore, understanding of it is at best elusive. In this thesis, we investigate this complex ecosystem from three dimensions. Firstly, we examine third-party trackers from a geographical perspective. We observe a non-uniform presence of local third-party trackers between regions and countries within regions, with some trackers focusing on specific regions and countries. Secondly, we focus on how trackers share user-specific identifiers (IDs). We identify user-specific IDs that we suspect are used to track users. We find a significant amount of ID-sharing practices across different organisations providing various service categories. Our observations reveal that ID-sharing happens at a large scale regardless of the user profile size and profile condition such as logged-in and logged-out. Finally, we quantify the effect of tracker-blockers, a popular option for the users to protect their privacy, on the page-load performance. The effect of such tools on the overall user browsing experience is questionable as the blockage of trackers can disrupt the general website loading process. The tracker-blockers we studied have a considerable negative effect on page-load performance. Unexpectedly, we find that even highly popular websites are negatively affected. This thesis points to significant gaps in our knowledge about the inner workings of this complex ecosystem. Moreover, it highlights some of the challenges that we face when attempting to preserve user's privacy by using tracker-blockers.

Contents

List of Figures

List of Tables

List of Abbreviations

ADNS Authoritative DNS.

CDNs Content Delivery Networks.

CSSs Cascading Style Sheets.

DNT Do Not Track.

NAI Network Advertising Initiative.

PLT Page-Load Time.

TACO Targeted Advertisement Cookie Opt-Out.

List of Publications

- Marjan Falahrastegar, Hamed Haddadi, Steve Uhlig, and Richard Mortier. The rise of panopticons: Examining region-specific third-party web tracking. In Alberto Dainotti, Anirban Mahanti, and Steve Uhlig, editors, Traffic Monitoring and Analysis, volume 8406 of Lecture Notes in Computer Science, pages 104–114. Springer Berlin Heidelberg, 2014.

- Marjan Falahrastegar, Hamed Haddadi, Steve Uhlig, and Richard Mortier. Tracking personal identifiers across the web, volume 9631 of Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pages 30–41. Springer Verlag, 2016.

- Rishab Nithyanand, Sheharbano Khattak, Mobin Javed, Narseo Vallina-Rodriguez, Marjan Falahrastegar, Julia E. Powles, Emiliano De Cristofaro, Hamed Haddadi, and Steven J. Murdoch. Ad-blocking and counter blocking: A slice of the arms race. USENIX Free and Open Communications on the Internet (USENIX FOCI), 2016.

Chapter 1

Introduction

The role of the Internet in everyday life evolves continuously. Interacting with Online Social Networks (OSNs), watching streamed videos and shopping online are all now daily activities in the lives of most citizens. In addition, Web interactions enabled by developments such as dynamic client-side interaction (e.g., Ajax [1]) and cloud-based services have led to significant changes in the Internet traffic [2] and website complexity [3].

One of the expanding family of new entrants in the Web are the third-party tracking services. They provide features such as advertising, analytics, OSN plug-ins and content hosting. Although some user interactions with these services may be conscious and explicit, e.g., sharing content or engaging with OSN plug-ins, most interactions users have with these services will not be explicit. Indeed, users may often be unaware of the presence of the third-party tracking services at all. Furthermore, these services can access user personal information and record user's online activities. For example, consider a scenario in which a user is visiting a web page about pregnancy on `www.nhs.uk` (the UK National Health Service); Figure 1.1 shows a sample of the HTTP requests that are sent from the user browser to `www.nhs.uk` and two third-parties: (1) `cloudfront.net` to serve content (e.g., images) and (2) `google-analytics.com` to analyse the traffic of the website. In this scenario, Amazon has access to the user's visited page through the `Referrer` HTTP header and Google explicitly records user's visited page and page

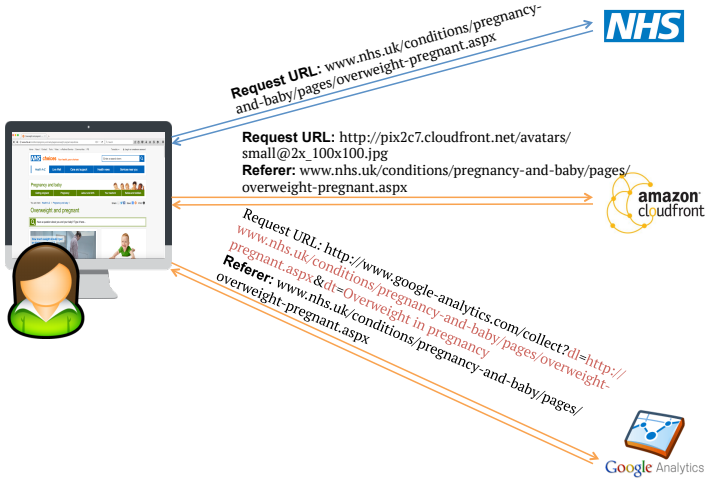


Figure 1.1: Example in which a user's visited web page on `www.nhs.uk` is exposed to third-party trackers: `cloudfront.net` and `google-analytics.com` via the `Referrer` HTTP header and the `dl` and `dt` URL parameters (coloured in red).

title via the `dl` and `dt` URL parameters. Therefore, a user's browsing activity is exposed to Amazon and Google without the user's awareness. The capability of third-party trackers to access user personal information has even been exploited for governmental surveillance purposes, e.g., the PRISMA surveillance program [4]. Additionally, third-party trackers are exploited by miscreants to spread malware across the Web [5].

Third-party tracking ecosystem has gained attention from academic, technical and legal communities, due to their association with privacy concerns and their increasing ubiquity. The activities of various communities help raise awareness about the potential risks caused by third-party trackers. There have been online projects that used third-party trackers to collect user information without their consent which failed due to the efforts made by technical and legal entities.[1] Therefore, we need further research to expose the inner workings of third-party tracking ecosystem to prevent future violations of user privacy. As a step towards this goal, this thesis aims to shed light on the prevalence of the third-party trackers and their potential effects on the end-users.

[1] For example, in a collaboration between BT and Phorm, a system was developed which collected user browsing behaviour to enable targeted advertisements. This project alarmed legal and technical communities which led to the suspension of the project.

1.1 Motivation

There are various studies focusing on identifying dominant players and their global distribution [6, 2, 7, 8]. They show that there is a limited number of corporations that are significantly dominating this field. Numerous works warn us about the capability of third-party trackers in collecting user personal information, this valuable asset of the Web economy [9, 10, 11, 12, 13, 14, 15]. Considering such a capability of the third-party trackers, it is expected that countries develop strategies to use their own local tracking services. However, we have limited insight into the local market of the third-party trackers. Indeed, how does the third-party tracking group look like besides the global players? Do non-dominant players have a prevalent presence? What countries have more local players? Are local players dominant? Are there some players targeting specific countries or geographic regions? In Chapter §4, we address these questions by analysing the geographical distribution of third-party trackers across 28 countries covering five geographical regions.

One of the practices of the user tracking is sharing user-specific identifiers (IDs). The parties involved in ID-sharing are in the position to merge their datasets corresponding to track the user whose user-specific identifier is shared. Current understanding about this practice is limited to a few works [16, 17]. They point to the importance of the ID-sharing mechanism by providing a valuable insight into the user browsing history's re-construction that can happen via ID-sharing. In Chapter §5, we present a study aimed to reveal the nature of ID-sharing parties across the Web.

As the tracking arms race continues, there are various efforts from different communities to provide protection mechanisms for the end-users. These mechanisms include the use of tracker-blocking plug-ins, opt-out cookies, private browsing and privacy-enhancing tracking. Tracker-blockers are currently the primary protection option for the informed users. Some studies investigate the effectiveness of tracker-blockers on user's privacy. However, the impact of tracker-blockers on the page-load performance is less-explored. This performance is observable by the end-users and thus has a vital role to keep them engaged with websites. Does blocking third-party trackers necessarily have a positive effect on the page-load performance? Do different tracker-blockers have similar effects on the page-load performance? How different categories of websites are being affected? In Chapter §6, we quantify the effect of two popular tracker-blockers, Ghostery and AdBlock Plus, to answer the raised questions.

1.2 Contributions

1. We reveal that **third-party trackers are beyond a set of global dominant players.** We observe a considerable number of local third-party trackers in various countries of our study of which some are dominant in their corresponding regions.

2. We reveal that **user-specific ID-sharing happens prevalently across organisations.** Furthermore, we uncover that the **ID-sharing practice happens independent of user's profile size and user's profile condition** (i.e., logged-in and logged-out).

3. We **quantify the effect of tracker-blockers on page-load performance**. We find that **tracker-blockers have different effects on the page-load performance** of websites.

1.3 Thesis Outline

We provide a brief background information about third-party trackers in Chapter §2. In Chapter §3, we present the state-of-the-art research investigating the aspects of third-party trackers that are within the scope of this thesis. We contribute to the existing knowledge in this field by revealing its global and local players across various countries in Chapter §4. Furthermore, in Chapter §5 we study the prevalence of a specific user tracking mechanism which employs user-specific ID-sharing, and we investigate the nature of the involved players. In Chapter §6 we analyse the effect of blocking trackers on websites in terms of page-load performance. We conclude this thesis in Chapter §7 by discussing the opportunities for future research directions.

Chapter 2

Third-Party Tracking Ecosystem

An ecosystem is a network of various elements interacting with each other and their environment [18]. Ecosystems have functions and purposes to accomplish [19]. In the context of third-party tracking, the purpose of the ecosystem is to deliver digital services and content in a way that makes money and keeps the involved businesses afloat [20]. This ecosystem has the following three main elements [21] (Figure 2.1):

- *Consumers*. End users that consume digital content and services provided by publishers,
- *Publishers (First-Parties)*. Creators and owners of content (e.g., The Guardian, Yahoo and Twitter) who provide consumers access to their content. Moreover, publishers allow third party elements (e.g., advertisers) to reach those consumers.
- *Third-Parties*. Publishers can use the services offered by third-parties to deliver content to their consumers. Some examples of the third-party services are targeted advertisements, analytic services and content hosting (see §2.1). Third-parties are able to track the users' interactions with the publishers.

The elements of an ecosystem are linked together via the interactions that happen between them. In the third-party tracking ecosystem, every interaction is in fact a flow of information. Consider a scenario in which *The Guardian* newspaper (a publisher element) requests targeted ad-

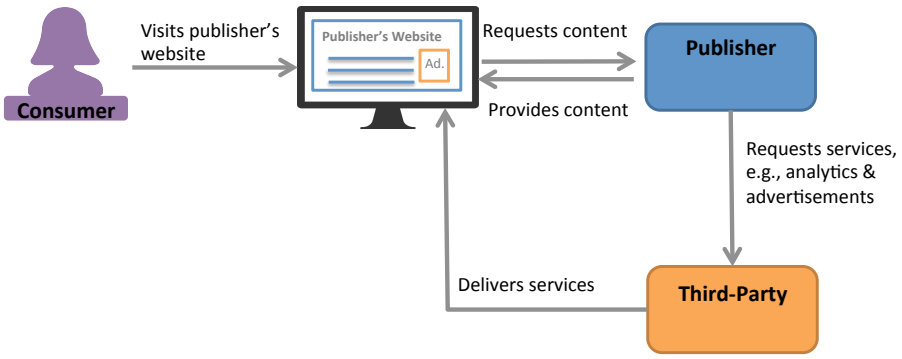


Figure 2.1: The main elements of the third-party tracking ecosystem

vertisements from a company like Double Click (a third-party element) for the visitor (consumer element) of its online website (`theguardian.com`). In this scenario, *The Guardian* sends various information to Double Click about the targeted page (e.g., content category) and the targeted visitor (e.g., unique user ID).

In this ecosystem, publishers typically offer free products and services that usually are not available offline, at least not for free. Meanwhile, publishers and third-parties can collect various information about the end users and track their activities across the Web. The collected information generates revenue (e.g., via targeted advertisement) that offsets the cost of providing content and services. However, this practice of user data collection raises serious privacy concerns (see §3.2).

We consider the third-party tracking ecosystem, a complex system; that is a system in which the elements can have heterogeneous types and interactions with each other [22]. These interactions are dynamic and are not present at all times which in turn can affect the behaviour and outcome of the system. For example, in the third-party tracking ecosystem, a single publisher website can collaborate with more than a hundred different third-party trackers [23]. Furthermore, the presence of some of these third-party trackers can depend on the visiting user-profile which in turn creates a dynamic set of collaborating elements in the ecosystem [17].

In this thesis, we take a multi-dimensional approach to explore the complex third-party tracking ecosystem. In the remainder of this section we describe the main dimensions and their significance within the ecosystem.

Geographic scope of players. The third-party tracking ecosystem contains publishers and third-party players from various countries across the world. While some of these players are globally dominant, some players target local markets and consumers. Various studies show us that US-based players are dominating the third-party tracking ecosystem. In Chapter §4 of this thesis, we investigate the presence of third-party trackers across popular websites of 27 countries. We identify local and global players across geographical regions.

Interaction between players. End users, publishers and third-party entities communicate with each other using standard protocols such as HTTP. Some of these communications are used to track the end users online activities. For instance, targeted advertising services try to find out users interests through monitoring users activities across the Web (see §2.1). Although different studies have addressed the user tracking mechanisms in the third-party tracking ecosystem (see §3.3), our understanding about the characteristics of the user tracking groups remains limited. In Chapter §5, we analyse the nature of players involved in tracking users and their potential intentions.

Privacy. There are various studies revealing the broad range of user personal information that is accessible to the third-party elements when visiting publishers' websites (see §3.2.2). Our observations in Chapter §5 also confirms the leakage of personal information such as health related information to the third-party entities. However, the end users are usually neither aware of this leakage nor have control over it. In response to users privacy concerns, various methods at the browser-side (e.g., browser plug-ins) are introduced to enable users observe the presence of tracking parties behind publisher websites (see §3.4). Additionally, users can control the presence of tracking parties (e.g., block tracking parties) using such methods. Although these methods aim to mitigate the effect of tracking parties, they may affect the general users' experience. In Chapter §6, we analyse the effect of blocking tracking parties on the performance of the publishers websites which reflects on users' experience.

Economy. The revenue model of various players in the third-party tracking ecosystem relies on collecting user personal information [24]. There are various studies estimating the financial value of personal information for the involved players and its subsequent impact on the economy of countries [24, 17, 25, 26]. However, the involved parties are usually secretive about the trade of user data because of the underlying privacy concerns and to maintain a competitive edge [27].

The economy dimension of third-party tracking ecosystem is out of the scope of this thesis. However, our geographic investigations of third-party trackers in Chapter §4 is useful to better understand the potential challenges for financial sectors. We further discuss these challenges in Section §7.2.

Data protection regulations. The third-party tracking ecosystem presents new challenges such as maintaining control over online personal data for the regulatory systems of countries. There are various works investigating and providing potential solutions for those challenges from the regulatory aspect [28, 29, 30]. In fact, some countries have already introduced *online* privacy regulations. However, their enforcement and effectiveness is of debate. Although the regulatory aspect is out of scope of this thesis, we briefly explain the regulations involving online privacy in some of the investigated countries and regions (see §3.2.1). Moreover, in Chapter §7 we discuss the challenges that countries encounter in dealing with the international presence of third-party trackers.

2.1 Third-Party Trackers

In this section, we describe our methodology for identifying third-party trackers. Moreover, we explain the main services that they provide and show how different third-parties can access and record users' information.

Identifying Third-Party Trackers. To identify a third-party tracker the domain name of a HTTP connection is compared with the domain name of the visited website. For example, `chartbeat.com` is a third-party tracker when `bbc.co.uk` is visited. However, this method is not accurate when one domain name is actually a DNS CNAME alias of another domain, for instance, `bbci.co.uk` and `bbc.co.uk`. To refine this domain-based approach, Krishnamurthy and Wills [6] introduced *Authoritative DNS (ADNS)* method. In this method, a third-party tracker of a visited site is the one whose ADNS server name is different from the ADNS server name of the browsed website. We can use Linux utility commands such as `nslookup` and `dig` to obtain the ADNS record of the domains. Figure 2.2 shows the ADNS records of `bbc.co.uk` and `bbci.co.uk` obtained using `nslookup` tool version 2.2.0. Comparing the *origin* field in the output of `nslookup` for `bbc.co.uk` and `bbci.co.uk` shows that their ADNS server name is the same. The implementation of the ADNS method is available in Appendix B.

Figure 2.2: The ADNS records of bbc.co.uk and bbci.co.uk (obtained using nslookup tool) indicate that their origin domain is the same.

2.1.1 Advertisement Entities

Advertisement entities are one of the most dominant categories of the third-party trackers. In fact, online advertising has become the main source of revenue for most websites [31]. Online advertisement services are able to serve advertisements that are customised based on visitor's demographics information, geographic location and interests.

The websites willing to use the advertisement services (i.e., publishers) need to embed the advertisement tracking programs. Figure 2.3 shows a sample tracking program of Google Ads service to be used by publisher websites. This program, as it is described by Google [32], collects statistics about user purchase, sign-up and page view. This tracking program includes a transparent image to categorise different pages of a website (e.g., purchase page) and some JavaScript software program to record user activities and report them back to Google.

There are various entities collaborating with each other to deliver online advertisements of which we describe the two main ones: Advertising Networks and Advertising Exchanges.

Advertising Networks. Advertising networks (ad networks) are intermediaries between advertisement sellers and websites that want to host advertisements (publishers). Ad networks choose targeted advertisements based on a targeting method (e.g., using visitor's location) specified by the publisher websites.

```
<noscript>
<img height=1 width=1 border=0
src="http://www.googleadservices.com/pagead
/conversion/1234567890/?value=1&label=Purchase&script=0">
<!----------------------------------------->
</noscript>
<script language="JavaScript"
src="http://www.googleadservices.com/pagead/conversion.js">
</script>
```

Figure 2.3: A sample tracking program provided by Google.

Advertising Exchanges. While ad networks facilitate the process of serving advertisement, a publisher may not be able to sell all of their advertisement slots through the ad networks, and for advertisers there is no guarantee that certain publishers will be selected. To solve this problem, an *ad exchange* entity is formed which enables selling advertisement slots via Real-Time Bidding (RTB) mechanism. In this mechanism (Figure 2.4) an advertising space is auctioned off in real time amongst different ad networks (or advertisement agencies) to put their bids based on the user's profile [31].

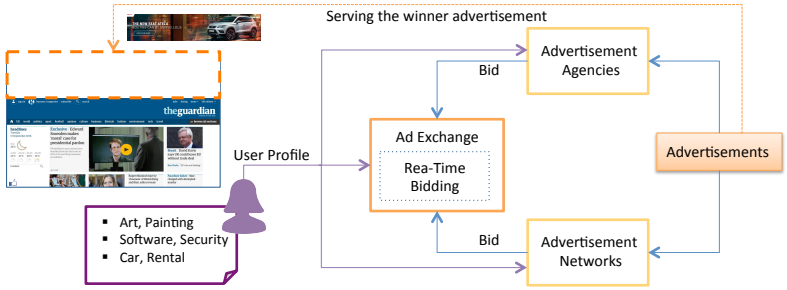


Figure 2.4: Ad exchanges sell the advertisement slots via Real-Time Bidding (RTB) mechanism. The involved parties bid based on the relevance of user profile to the advertisements that they offer.

2.1.2 Web Analytics Services

Web analytics services enable websites to better understand their audiences. These services provide various statistics such as number of unique visitors, their demographics and their browsers specifications. Similar to the advertisement entities, analytics services provide tracking software programs to be embedded in the publisher websites. For example, each time a user visits `cnn.com` which uses the QuantCast analytics service, a HTTP request is sent to QuantCast shown in Figure 2.5. The parameters in the requested URL reports information such as user visit via the *url* parameter, category of user interest via *title* parameter and user screen resolution via

the *sr* parameter.

RequestURL:	`http://pixel.quantserve.com/pixel;`*a*`:p-D1yc5zQgjmqr5;`*url*`:` `//www.cnn.com/politics,`*title*`:Politics-PoliticalNews,` `AnalysisandOpinion;`*sr*`:1280x800x24`
Referrer:	`http://edition.cnn.com/`

Figure 2.5: A sample HTTP request sent from `cnn.com` to QuantCast analytic service reporting various information about user via URL parameters (shown in *italic*).

These services have different policies regarding data collection and data sharing. For instance, Adobe guarantees by contract not to access or use the collected data [33]. Google provides an opt-out option for the publishers by setting a parameter in Google Analytics JavaScript software program [34]. Moreover, Google provides a browser add-on for users who wish to disable tracking [35].

2.1.3 Online Social Networks

Online Social Networks (OSNs) differ from the other categories of third-party trackers due to their dual role as being first-party (directly visited by user) and third-party simultaneously. The OSN providers have access to two sources of user data, one that is directly provided by their members and the other one that is indirectly collected (as third-parties) by tracking users. For instance, Facebook is considered as a third-party tracker when other websites use its services such as its social plug-ins e.g., *like* button and Facebook Audience Network (an advertisement service) [36]. When a website such as `booking.com` uses an OSN service like the Facebook advertisement service, detailed information of visitor's query on `booking.com` is reported to Facebook. Figure 2.6 shows a sample HTTP request sent from `booking.com` to Facebook. The parameters in the requested URL such as *cd[action]* and *cd[city]* show the user's query. In this scenario, if the user is subscribed to Facebook, the collected user data via `booking.com` can be aggregated with the user's Facebook profile information using Facebook user-specific identifier.

RequestURL:	`https://www.facebook.com/tr?`*id*`=405133399621612&`*cd[action]* `=searchresults&`*cd[city]*`=Sydney&`*cd[checkin_date]*`=2016-12-24.` `..`
Referrer:	`http://booking.com/searchresults.en-gb.html?`

Figure 2.6: Facebook as a third-party tracker for `booking.com` can record user information that is included in the URL parameters (shown in *italic*).

2.1.4 Hosting Services and Content Providers

Hosting Services enable websites to reduce load from their servers by transferring the content to other servers. For instance, Content Delivery Networks (CDNs) replicate content over various servers located in different geographic locations, and serve the content from the close-by server to the user. For example, a website like `foxnews.com` uses Akamai CDN service to host some of its content. In this case, Akamai can record user's visit via URL parameters (Figure 2.7). In addition to delegating hosting of content to the third-party trackers, many websites use content and applications provided by third-parties such as weather forecast, map, News feeds, *etc.*

RequestURL:	`https://vvq6xvgu5f-a.akamaihd.net/z.gif?`***d***`=foxnews.com;agency-probes-whether-california-dem-party-funneled-illicit-oil-donations-to-governor;_page&`***t***`=1474928411108&`***i***`=FOXNEWSCONTENT`
Referrer:	`http://www.foxnews.com/politics/2016/09/26/agency-probes-whether-california-dem-party-funneled-illicit-oil-donations-to-governor.html`

Figure 2.7: Akamai CDN hosts some content of `foxnews.com`. Akamai includes the visited page by user in the URL parameters (shown in *italic*).

2.2 Summary

In this chapter we describe the main categories of services provided by third-party trackers. Some categories such as advertisement entities include various players e.g., advertising networks, advertising exchanges and advertisement providers. Analytics services provide statistical information about websites' visitors. OSN services enable websites to increase their visitors interactions. Hosting services serve contents of websites and application providers enhance websites with commonly used applications such as map and weather forecast.

We show how user personal information is accessed within each category. While the third-party trackers mostly have indirect access to user personal information, those in OSN category are privileged with direct access to the user profile information of their members.

The distribution of third-party trackers varies across different categories of services. For example, the advertisement entities and the hosting services are more dominant in comparison with the other categories. In Chapter §4, we give insight about how third-parties are distributed across categories of services. Moreover, third-party trackers can collaborate with each other to build a

more accurate user model. In Chapter §5 we investigate the collaboration between third-party trackers of different categories.

Chapter 3

Literature Study

The Third-party tracking ecosystem has been studied from various perspectives. Some studies elaborate on the evolution of third-party trackers over time [6, 2, 7, 8]. Other studies inform us about the potential privacy violation risk of third-party trackers by investigating their access to user personal information [9, 10, 11, 12, 13, 14, 15] and analysing their techniques for tracking user activities across the Web [37, 17, 38, 39, 40, 41, 42]. Some studies focus on the techniques to mitigate user privacy violation and protect users from being tracked by providing browser-based solutions [43, 44, 45, 13, 46, 47, 48, 49] and privacy preserving user tracking methods [50, 51, 52, 53]. In this chapter, we review the state-of-the-art studies covering the aforementioned aspects of the third-party tracking ecosystem.

3.1 Evolution of the Third-party Trackers

The evolution of the Web, in particular, third-party trackers has been studied in various works. A number of longitudinal studies illustrate the gradual penetration of third-party trackers across the Web. Krishnamurthy and Wills [6] measure the penetration of the third-party trackers amongst 1200 English-language popular websites across different categories of Alexa during three years, from 2005 until 2008. They observe an increasing penetration of the third-party trackers amongst the studied websites from 40% in 2005 to 70% in 2008. Beside individual third-party trackers,

they investigate the presence of dominant companies in this industry. They manually identify the acquisition of some third-party companies by prominent companies including AOL, Google, Microsoft, Omniture, ValueClick and Yahoo. They report that the presence of Google's third-party tracking services amongst the websites reaches almost 60% in 2008 from only 8% in 2005. After the Google family, Omniture and Microsoft have the highest growth during the period of their study although much less than Google. They take one of the first steps to provide insight into the growing market of the third-party trackers. We show further extension of tracking services in these companies and emergence of new dominant companies providing tracking services (§4.4).

Another longitudinal study of the Web ecosystem is the one done by Ihm *et al.* [2]. In that study the changes of the Web traffic during four years, from 2006 until 2010, across four countries (US, Brazil, China and France) were examined. Their dataset includes their local university's network traffic collected through a globally distributed proxy on top of PlanetLab calling CoDeeN [54]. They examine the traffic share of top 50 advertising networks and analytics sites across the mentioned countries. Their findings show a consistent growth of such traffic across the investigated countries with the largest growth in Brazil, in which the advertising networks traffic acquired 12% of the traffic in 2010 in comparison with 1% in 2006.

Simpson *et al.* [7] present the longest longitudinal study of the 2third-party trackers during 10 years, from 1996 to 2016. They studied Alexa top-500 websites using the Internet Archive's Wayback Machine. Wayback Machine hosts an archived version of websites since 1996. The majority of popular third-party trackers are represented via the Wayback Machine, whereas the presence of some is missed due to technical problems when websites are archived. However, their investigation shows that the Wayback Machine's data is reliable to study the trends of third-party trackers changes over time. In general, they observe a sharp rise in the number of third-party trackers creating cross-domain cookies (§6.2) from 2012 to 2016. Additionally, they find that the number of third-party trackers forcing users to visit their websites (e.g., through pop-up windows) are at peak in mid-2000, whereas this number is considerably reduced in the recent years.

Castelluccia *et al.* [8] examine the penetration of US-based third-party trackers on the popular websites of 37 countries in 2013. Their examinations show the dominant presence of US-based trackers amongst almost all studied countries. Russia appears as an exception with 49% local trackers which dominates 39% US-based ones. We note that in this study AdBlockPlus and

Ghostery Firefox plug-ins were used to identify third-party trackers. Therefore, the identified third-parties are limited to those stored in the database of Ghostery and AdBlockPlus (see §6.2). However, their observation uncovers the global demand for using third-party tracking services.

The current studies have revealed the speedy growth of third-party tracking services. The big and mainly US-based corporations have taken the lead in providing such services. Moreover, the access of these services to user's data, as one of the most valuable capitals of the Web, is a great motive for governments and countries to have their local players in this industry. In Chapter §4 we study the regional, local and less-known players to show that, indeed, local and regional players have a strong presence across popular websites of different countries.

3.2 Privacy Concerns

There are various studies examining the accessibility of personally identifiable information to third-party trackers. These studies shed light on the potential risk of privacy violation by third-party trackers. Therefore, it is necessary to further explore third-party trackers.

3.2.1 Personally Identifiable Information (PII) and User Data Protection.

Data privacy law defines Personally Identifiable Information (PII) and provide a structure to protect user's data. However, the strictness of data privacy law varies across different countries. Moreover, some countries have introduced *ePrivacy* laws focusing on user's privacy on the Web. US privacy law defines PII [55] as "any information which can be used to distinguish or trace an individual's identity, such as their name, social security number, biometric records, *etc*. alone, or when combined with other personal or identifying information which is linked or linkable to a specific individual, such as data and place of birth, mother's maiden name, *etc.*.". Some examples that may be considered as PII (Table 3.1) are [56]: name (e.g., full name and maiden name), personal identification number (e.g., social security number), contact information (e.g., street address, email address and phone number) and device information (e.g., IP and MAC address). Some pieces of PII can be more harmful if exposed, for example the leakage of an individual's Social Security Number, medical history and financial account can be more harmful than an individual's phone number.

In the US, the enforcement of regulation varies across different states and industries. Additionally, some states have their own privacy laws of which California has one of the best user privacy and data protection laws [57]. Under California law [58], ”any company that tracks any personally identifiable information about consumers must disclose in its privacy policy whether the company honours any Do-Not-Track method or provides users a way to opt-out of such tracking.”. The same law also requires website operators ”to disclose in their privacy policy whether any third parties may collect any personally identifiable information about consumers on their website and across other third party websites.”.

The EU is considered to have stricter privacy laws than the US. The EU Directive 95/46/EC [59] defines personal data (a term similar to PII) as ”any information relating to an identified or identifiable natural person (”data subject”); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity.”.

The EU regulatory framework has a clear and comprehensive set of rules for personal data protection. This means that all businesses located in any members of the European Union should comply with such regulations. According to the Directive 2002/58/EC and its amendment 2009/136/EC, known as ePrivacy Directive, websites that are using cookies or other technologies to collect user data should clearly inform users about such process and ask for their opt-in consent as soon as website is loaded on user’s machine.

In the EU, the implementation and enforcement of data protection legislation can vary across members of the European Union. For example, in Germany there is no requirement for active opt-in consent, e.g., by clicking on a pop-up window containing opt-in consent message. While in countries like UK, Netherlands and France having user opt-in consent is mandatory. Similarly, in Australia, any entity that collects personal information should notify user about that. However, the notification can be provided after the actual data collection [60].

We note that the EU’s strong data protection rules make any non-EU country, willing to do business in the EU, to define adequate levels of data privacy practice in-line with the EU standard. For example, Argentina’s data protection law has been updated to provide broader protection of personal data following Spain’s data protection law [61]. Turkey is another example that has provided new data protection law in 2016 as an attempt to harmonise with the EU [62].

Contact info.	Home address, City, Phone and Mobile number, Email address
Device info.	IP address, MAC address, Mobile Unique Identifier
Demographic info.	Age, Gender
Sensitive info.	Social Security Number, Medical history

Table 3.1: Samples of Personally Identifiable Information (PII).

In Asia also some countries are adapting online privacy regulations. South Korea has one of the strongest data protection laws in Asia [57]. Under South Korea's IT Network Act, personal information is defined as "information pertaining to a living individual, which contains information identifying a specific person with a name, a national identification number, or similar in a form of code, letter, voice, sound, image, or any other form." [63]. Moreover, using cookies is also regulated and should provide opt-out option for the users. In China, under the cyber security Law [64], personal information is defined as "including all kinds of information, recorded electronically or through other means, that taken alone or together with other information, is sufficient to identify a natural person's identity, including, but not limited to, the natural persons' full name, date of birth, identification numbers, personal biometric information, addresses, telephone numbers and so forth". In contrast with the EU, there is no particular requirements for cookies in the existing regulations. Similar to the US, China currently lacks a centralised enforcement mechanism for data protection and there is no single data protection authority or any other state agency established to monitor the protection of personal data [64].

3.2.2 Accessible Personally Identifiable Information

Krishnamurthy *et al.* in [9] examine the access of third-party trackers to PII. They study over 100 popular websites across various categories such as news, shopping and relationship. They focus on a subset of websites that enable users to create an account. After creating accounts and navigating through the websites, they search the inserted information during account creation amongst the recorded HTTP requests. They find the presence of *fullname* and *email* information amongst the requests sent to the third-party trackers from 20 websites. There are various studies showing that unique *user-identifiers* (created by websites to identify their users) are being passed to third-party trackers [10]. Chabaan *et al.* in [11] report that 18% of the applications embedded in Facebook (e.g., game applications) transmit the Facebook user-identifier to the third-party trackers. Third-party trackers accessibility to personal user information is even, sadly, extended

to *sensitive information* (e.g., health related topics). Wills and Tatar [12] show that some advertisements served by the Google Ad Network and Facebook Ads are targeted based on sensitive searched topics by the user. For instance, a user is served an anti-depression treatment advertisement after searching for depression. Additionally, the presence of sensitive searched-terms in the URL parameters of the requests sent to the third-parties is reported in [13, 9]. Unfortunately, we also observe access of third-party tracking groups to sensitive user information (see Chapter §5).

RequestURL:	`ads,mobclix.com?`*i*`=xxxxxxxx-xxxx-;`*u*`=IPHONE-UDID;`
User-Agent:	`Wattpad/1.6.1CFNetwork/459`

Figure 3.1: Mobile unique device-identifier is transferred from `Wattpad` mobile application to `mobclix.com` via *i* URL parameters.

Third-party trackers can access a new class of PII including precise *user location* and *device-identifier* when users access the Web via mobile devices. Krishnamurthy and Wills [14] examine the leakage of these new types of information when users use mobile versions of Online Social Network (OSN) websites and also OSN mobile applications. They find that six out of 13 studied mobile OSN services transfer device-identifiers to the third-party trackers. Figure 3.1 shows an example in which `Wattpad` transfers a unique device-identifier to `mobclix.com`. Additionally, they find two (out of 13) mobile OSN services that transfer precise user locations to the examined third-party trackers. Figure 3.2 from [14] shows an example in which the precise user location (longitude and latitude specifications) are transferred from `Buzzd` application to a third-party (`pinchmedia.com`).

RequestURL:	`http://beacon.pinchmedia.com?`*lat*`:20.00;`*lon*`:-70.00`
User-Agent:	`buzzd/2.2.0CFNetwork/459`

Figure 3.2: The precise user location (latitude and longitude information) is transferred from Buzzd mobile application to `pinchmedia.com` via *lat* and *lon* URL parameters.

User Data Aggregation. Third-party trackers can expand their knowledge about users by aggregating their dataset with anonymised datasets. In fact, information such as commercial transaction, web browsing behaviour and search history can distinguish users from each other. Narayanan and Shmatikov [15] examine the identification of the users subscribed on different social networks. They showed that 30% of users who are members of both Flicker and Twitter can be identified from anonymous Twitter graph when there is a 15% overlap with the Flicker dataset. Additionally, third-party trackers can expand their user dataset by aggregating it with publicly

available personal information datasets. For instance, RapLeaf.com, one of the leading data trading companies, provides name and other personal information if their client already have a person's email address [65].

3.3 User Tracking Mechanisms

Third-party trackers use various techniques to keep track of users browsing activities. Roesner *et al.* [37] report that the majority of the third-party trackers of Alexa top-500 websites create their own cookies and assign their own domain name to the created cookies. Figure 3.3 shows an example of a cookie that belongs to doubleclick.net that stores a user-identifier. In this example, doubleclick.net can read the stored user-identifier in that cookie from any website in which doubleclick.net is embedded. Hence, doubleclick.net can track users through various websites. Arnold Roosendaal [66] reports that when a user visits a website em-

RequestURL:	https://adx.g.doubleclick.net/pagead/adview
Set-Cookie:	*id*:35c192bce000b1;*domain*:doubleclick.net;

Figure 3.3: A sample HTTP cookie set by DoubleClick storing a user-identifier.

bedding the Facebook social plug-ins (e.g., Like button), Facebook receives the user-identifier via its cookie. Krishnamurthy *et al.* [9] report cases in which one third-party tracker includes its cookie value in a request sending to another third-party tracker i.e., *cookie syncing*. Olejnik *et al.* [17] study cookie syncing as part of an in-depth investigation of Real-Time Bidding (§2.1.1) characteristics. They observe the presence of over 100 cookie syncing across the Alexa top-100 websites. Acar *et al.* [38] further investigate the effect of cookie blocking on cookie syncing, in particular those cookies storing user-identifiers. They introduce a method for detecting the user-identifiers in cookies. They observe that cookie blocking decreases the amount of cookie syncing and the number of involved parties. However, they reveal an important point that such a decrease does not affect the overall access of domains to user data if these parties merge their datasets in back-end. These studies highlight the presence and importance of cookie syncing practice across the Web. In Chapter §5, we complement these studies by focusing on the nature of user-specific ID-sharing players and their relation with user information.

While a cookie is the main location used by third-party trackers to store user tracking information, other available storage at the end-user side have been exploited as a backup for cookies.

Soltani *et al.* [39] report an unexpected use of *Flash cookies* (Local Shared Object) [67] as a mirror of HTTP cookies. Flash cookies have some advantages over HTTP cookies such as a larger storage capacity and not having an expiry date. Soltani *et al.* identify 31 websites out of the top-100 Quantcas websites having at least one Flash cookie corresponding to one of their HTTP cookies. Mcdonald *et al.* [40] report the regeneration of HTTP cookies using backup content in Flash cookies after user removal. Acar *et al.* [16] report 107 websites amongst the top-10K examined websites that are using Flash cookies. Other unconventional storage locations can be used by third-party trackers, including `ETag` (Entity Tag) and `Last-Modified` HTTP fields [68, 69]. For example, `hulu.com` (a video streaming website) is reported to use the ETag field as a backup for user information stored in HTTP cookies [69]. Samy Kamkar [70] introduces the *ever-cookies* API to replicate data into various storage locations including Flash cookie, ETags and browser database to generate persistent cookies.

Some third-party trackers have recently turned to fingerprinting mechanisms to identify and track users. Mowery *et al.* [71] introduce a new browser fingerprinting method based on the Canvas API in HTML 5, called *Canvas fingerprinting*. This API enables a browser to generate images on-the-fly. They find that the specification of an image generated by the Canvas API is unique for each machine due to its dependencies on some system specifications including the operating system, font library, graphics card, graphics driver, display resolution and the browser. This may be due to the differences in system fonts, API implementations and the physical display. Acar *et al.* [16] reveal that 5.5% of Alexa top 100K websites run canvas fingerprinting scripts on their websites. Surprisingly, 95% of the discovered fingerprinting scripts belong to a single third-party tracker, namely `addthis.com`. Eckersley *et al.* [41] investigate the effectiveness of fingerprinting based on the browser specifications (e.g., version, plug-ins, and *etc.*) that are inferred from HTTP requests. They find 94% of browsers (with enabled Flash or Java Virtual Machine) exhibiting unique fingerprints. Olejnik *et al.* [42] study the feasibility of fingerprinting based on user browser history. They find that 69% (196 out of 284) of their participants have unique browsing history. Nikiforakis *et al.* [72] investigate the different implementations of fingerprinting libraries. They report a frequent use of Flash and Internet Explorer's specific properties such as `navigator.securityPolicy` and `navigator.systemLanguage` in the fingerprinting libraries.

Browser-side	Tracker-blocking browser plug-ins, Opt-out Cookies, Do Not Track header, Private Browsing
Privacy Preserving User Tracking	Privad, RePrive, Adnostic

Table 3.2: Methods to mitigate the risk of personal user data collection.

3.4 Protecting Users

To protect users from invasive user tracking and mitigate the risk of personal user data collection, various methods have been introduced at the browser-side, such as blocking trackers, private browsing, opt-out and DoNotTrack signals. Moreover, various proposals have been given to modify the interaction model of third-party trackers accordingly to preserve user privacy (Table 3.2).

3.4.1 Blocking Third-Party Trackers

One of the approaches to restrict user tracking is to prevent websites from sending requests to third-party trackers. 2There are various tools to block third-party trackers using predefined tracker-blocking lists.The analysis of 12 blocking tools done by Mayer [43] shows that understanding and configuring such tools can be challenging for the users. For instance, some tools such as Ghostery [73] do not block the trackers by default, unless the users configure the corresponding settings, however, many users do not feel the need for such configurations [74].

Some tracker-blockers such as *Milk* [44] and *PrivacyBadger* [45] interrupt third-party tracking mechanisms instead of blocking them. For instance, Milk stops third-parties from creating cross-site cookies, and enforce them to create a new cookie for each website they are embedded in (§3.3). Additionally, some tools are aimed at raising user awareness about the presence of third-party trackers. For instance, *NoTrace* [13] notifies users about their personal information that are passed to the third-parties while they are visiting websites. This tool uses regular expressions to detect pre-defined fields of interest in the URL such as email, name, *etc*. Despite the shortcomings of these tools, using tracker-blocker tools has become popular with an estimated number of 198 million active users [75]. In Chapter §6, we analyse how these tools affect websites in terms of page-load performance. Our analysis reflects also on the user's experience.

Opt-out Cookies. The Network Advertising Initiative (NAI) which is in charge of online

advertising regulation in United States, mandates targeted advertisers to provide users the choice of opting out from their services. NAI implemented this principle using opt-out cookies [47] to notify the ad network about user's preferences. There are several concerns regarding this method: users need to apply this setting for each browser separately; they need to create a new opt-out cookie for each ad network; opt-out cookies can be removed mistakenly when users clear their cookies; third-party trackers can set or delete opt-out cookies. Some of these problems have been mitigated using browser extensions. For example, Targeted Advertisement Cookie Opt-Out (TACO) [76] is a Firefox plug-in to simplify the usage of Opt-out cookies. It automatically sets opt-out cookie for the ad trackers who are the member of NAI. However, Komanduri *et al.* [46] report various cases amongst NAI member that do not follow NAI principles in regards to the opt-out option.

Do Not Track. *Do Not Track (DNT)* is a proposed HTTP header field to give a signal to web sites to disable their tracking. This proposal is easily implementable from the browser-side by appending DNT=1 field to the HTTP outgoing request; Chrome, IE, Firefox and Safari have already implemented it. However, the action that needs to be taken by the receiver-side of this header (e.g., publisher websites and third-party trackers) is not clearly defined. In comparison with opt-out cookies, this method is robust and can be easily added to browsers, yet there is no technical nor legislation enforcement behind it, which sadly left it as a policy only [48].

3.4.2 Private Browsing Mode

Major browsers such as Chrome, Firefox, IE and Safari include the *private* browsing mode feature. The goal of private mode is to enable users to browse across the Web without leaving any trace of their activities on their machine. To achieve this goal, user browsing activities during private mode should not be accessible in the standard mode (browser is not in the private mode) and vice versa. Moreover, user activities across different sessions in private mode should not be linkable. Aggarwal *et al.* [49] investigate the implementation of private browsing across four popular browsers. They report that Firefox, Chrome and IE handle the secure transition between standard and private mode, whereas Safari leaves storage components of the standard mode available in the private mode which is against the aforementioned goal of private browsing mode.

3.4.3 Privacy Preserving User Tracking

Privacy Preserving user tracking techniques, in contrast to the previous methods, enable third-party trackers to collect information about users while preserving their privacy. For instance, *Privad* [50] is a privacy preserving architecture for targeted advertisement. In this system the user profile is built at the client-side. Users subscribe to a specific category of advertisements and chooses the most relevant ads based on their interest. The client software reports the visited ads to the Ad Network (e.g., DoubleClick) via a proxy in-between to hide user's identity. Privad is an ideal architecture which blocks any user information leakage (such as user interests, user ad view). However it needs fundamental changes to the current business model. Similar to Privad, *Adnostic* [51] relies on the client-side profiling and local advertisement selection. It uses an intermediary component as a *trusted third-party* which is in charge of encrypting and decrypting the user interaction.

In contrast to [50, 51], *RePrive* [52] is a general-purpose architecture for controlling the transmission of personal information through the browser. It keeps user information and the profiling process at the browser. Moreover, RePrive enables third-party applications (e.g., search engines and OSNs) to run their own data mining on the browser via a browser extension. The third-party extensions should comply with security policies defined by RePrive such as informing users about access to their data. RePrive grants permissions to users to decide about releasing their information. While this architecture can support various applications, its practicality is dependent on the service providers to accept the user profile maintained by RePrive. Additionally, it is confusing and inconvenient for the users to be asked for their consent in any communication with the third-party trackers. Bilenko and Richardson in [53] propose a client-side approach for keyword-search advertising. In this approach the computation for building the user profile is done by advertisement entities. The user profile is stored at the client side and the company is trusted to delete the profile from their server. This approach requires less changes in the current system, however, the question of how to guarantee the compliance of the involved parties remains challenging.

We note that technical and research communities have provided various guidelines to make system designers and implementers aware of privacy-related design choices. One of these communities is Internet Engineering Task Force (IETF), the organisation that defines standard Internet op-

erating protocols such as TCP/IP [77]. IETF has issued several guidelines such as RFC-6973 [78] which lists different kinds of privacy threats, mitigations for those threats and a check-list of questions for identifying and addressing privacy issues. Additionally, in RFC-7258 [79] the threat of pervasive user monitoring and the necessity to consider such a threat in the technical design of both new and existing protocols are discussed.

3.5 Summary

In this chapter we presented the state-of-the-art research investigating third-party tracking ecosystem. We discussed how third-party trackers attempt to collect personally identifiable information, exchanging with other parties and the approaches to avoid these actions. In the remaining chapters we extend the existing studies by further exploring third-parties, their geographical presence, the nature of user tracking groups and the effect of blocking trackers on page-load performance of websites.

Chapter 4

Anatomy of the Third-Party Trackers [1]

4.1 Introduction

Third-party trackers have become an uninvited guest in our online browsing activities. Third-party tracking ecosystem is evolving in scope and complexity [7, 6, 3]. Third-party trackers combine multiple tracking methods [7, 37] and collaborate with various other trackers [81] to expand their capability in collecting personal data from users such as demographic information, interests and locations.

Now-a-days personal data is an important asset for the digital economy similar to how the crude oil is for the industrial economy [8]. Hence, the capability of third-party trackers to collect personal information made them popular for the digital economy. For example, Facebook made over $3 billion revenue from targeted advertisements in the second quarter of 2015 and only UK advertisements are estimated to have contributed £5 billion to Google's sale in 2015 [82].

While these services are now vital for the digital economy, they have triggered plenty of debate regarding the user privacy violation issues as the end-user is usually unaware of such services and has no control over them. Moreover, these services can be exploited by governments for citizen surveillance. For example, Google analytics service has been misused by the National Security

[1] This study has been published in the proceedings of the 6th Workshop on Traffic Monitoring and Analysis (TMA), 2014, UK [80]

Agency (NSA) in the USA to identify individuals [4].

Previous works [8, 6], As reviewed in Section §3.1, [8, 6] report that a small number of international corporations are heavily dominant across the globe including corporations such as Google, Yahoo, Microsoft and AOL. They observe that the third-party tracking services of these corporations are growing in terms of the number of their services and their penetration across websites. Specifically the presence of Google third-party services has grown 52% during three years (years 2005 to 2008). We expect that organisations such as Google are still dominating due to the increasing variety of their services. However, we expect the rise of other global and local players due to the growing popularity of third-party tracking services, in particular advertisement related services.

US-based third-party trackers have taken the lead in this industry. Considering the different applications and benefits of third-party trackers (see §2), it is expected that countries may develop their own local tracking services. Some countries like Russia have already expanded their local third-party tracking services [8] (see §3.1). We expect that more countries develop their local tracking services, specifically those with higher level of Web penetration. We believe that understanding the geographical distribution of third-party trackers, particularly local ones, helps to get better insight of this growing industry. In this chapter, we expose the third-party tracking industry across various countries and regions to understand:

- The correlation between the number of third-party trackers and the Web penetration in the countries.
- The role of third-party trackers in different countries.
- The distribution of the local and overseas third-party trackers across countries and geographic regions.
- The distribution of major and minor players in various geographic regions.

In the following sections we first describe our data collection method (§4.2). We then analyse the presence of third-party trackers on both per-country (§4.3) and regional (§4.4) basis. We summarise our findings in §4.5.

Figure 4.1: The data collection procedure is performed in the depicted three steps.

4.2 Data Collection

We develop the find-tracker Firefox browser plug-in to record all the HTTP requests and responses passing through the browser (the plug-in is available online [83]). Our plug-in uses the Firefox API [84] to capture HTTP headers such as `Host`, `Referer` and `Cookie`. Figure 4.2 shows a sample of an HTTP request sent from `www.nhs.uk` (`referer` header) to `webtrendslive.com` (`Host` header) which includes a `cookie` set by `webtrendslive.com`. Additionally, we build a Python software program that accepts a list of URLs as its input and subsequently opens up the browser to visit each of these URLs one at a time. Moreover, the Python program removes the existing Firefox profile and creates a new profile before visiting each website (This software program is available in Appendix A). This program together with our browser plug-in, enable us to collect the HTTP requests and responses passing through the browser when a website is visited.

Figure 4.1 shows the data collection procedure in our experiment. We note that this procedure is performed automatically by the aforementioned Python program. First, the find-tracker plug-in is installed on the Firefox browser. Second, a new Firefox profile is created. Afterwards, the landing page of the Alexa top-500 popular websites in each country listed in Table 4.1 is browsed. While the landing page is open on the Firefox browser, the plug-in records all the HTTP requests sent by the browsed website. Every 60 seconds, the next URL is retrieved from the list and its landing page is visited as explained above.

We use PlanetLab's infrastructure nodes to carry out our experiment. We run our experiment in 28 different countries to gain access across the globe. All the nodes are identically set-up by installing GNOME desktop environment and Firefox. To improve our coverage in the Middle-East, we also run our experiment on a computer located in Qatar. Further, the paucity of PlanetLab nodes in Africa is coupled with the failure of our scripts to complete successfully on the few nodes in Africa, we do not present data for that region. All our data are obtained between 28

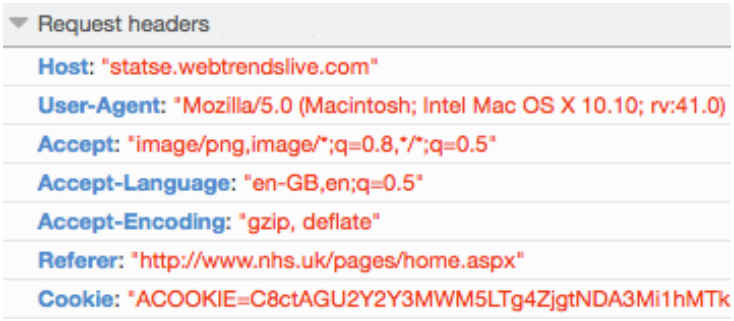


Figure 4.2: A sample of HTTP request sent from `www.nhs.uk` (`referer` field) to webtrendslive.com (`Host` field) which includes a `cookie` set by webtrendslive.com.

March 2014–28 April 2014.

Region	**Country**
North America	Canada, US
South America	Argentina, Brazil, Ecuador
Europe	Belgium, France, Germany, Greece, Hungary, Italy, Netherlands, Norway, Russia, Slovenia, Sweden, United Kingdom
East Asia	China, Hong Kong, Japan, Korea, Taiwan
Middle East	Israel, Jordan, Qatar, Turkey
Oceania	Australia, New Zealand

Table 4.1: The countries for which we collected data and their assigned region.

We identify third-party trackers using the ADNS method that we discussed earlier in Chapter §2. In visiting the Alexa top-500 websites in 28 countries from different regions of the world, we visit a total of 6,497 unique websites and identify 6,817 third-party trackers. We observe the presence of third-party trackers on over 80% of the visited websites. Qatar (814), Korea (769) and Hong Kong (726) are the top three countries in terms of number of third-party trackers, while the United Kingdom (397), Jordan (330) and Belgium (274) are the bottom three. We group countries into six geographical *regions*: North America, South America, Europe, East Asia, Oceania and the Middle East. Table 4.1 shows the investigated countries and the regions to which they belong. The highest numbers of third-party trackers is seen in Europe (3378) and East Asia (2009). Normalising by the number of countries in each region, North America, Oceania and the Middle East are the regions with the highest average numbers of third-party services per country.

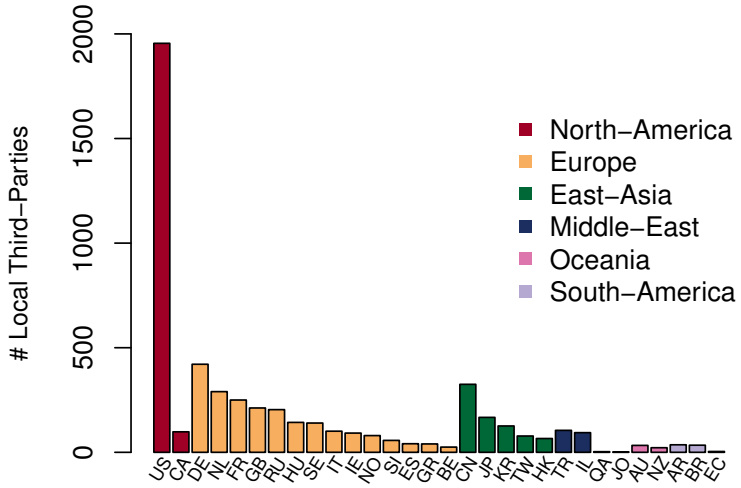


Figure 4.3: The strength of countries in terms of number of local third-party services.

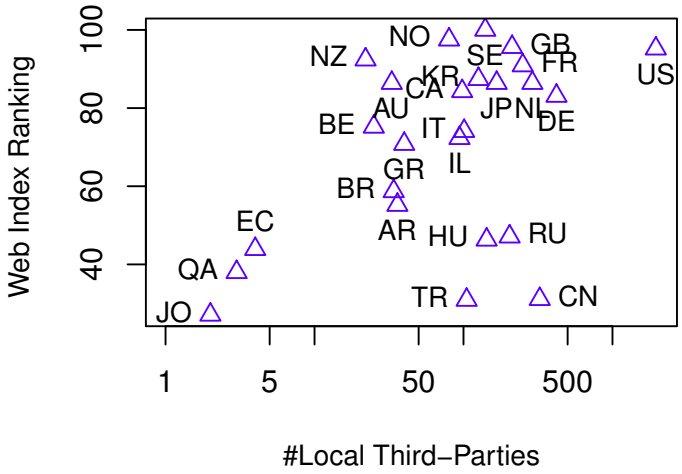


Figure 4.4: Web Index ranking against locally hosted third-parties per country.

4.3 Per-Country Analysis

In this section, we examine the presence of third-party trackers within and across the countries of our study. We begin our analysis by counting the number of third-parties which are physically hosted in the studied countries. We rely on a `geoiplookup` utility tool to determine the country in which each observed third-party resides. The `geoiplookup` tool looks up for a given host-name or IP address in the online available GeoIP databases such as those from MaxMind [85]. Our results in Figure 4.3 show substantial variations in number of locally hosted third-parties in the countries across each region. For example, in the Middle East region Turkey and Israel have many more local third-parties than other countries in that region. Overall, we find that US, Germany and China have the highest number of locally hosted third-parties.

Is there a correlation between level of Web penetration in a country, and the number of third-parties hosted in that country? To answer this question, we use the Web Index[2] that is provided by the WWW Foundation. The index, first released in 2012 and updated in 2013, measures the contribution of the Web in 81 countries using four factors: "Universal Openness" for communication infrastructure, "Freedom and Openness" for citizen rights of information, opinion and online privacy, "Relevant Content" for accessibility of relevant information based on gender and language, "Empowerment" for impact of the Web on society, economy and politics. Figure 4.4 presents the scatter-plot for Web Index ranking against locally hosted third-parties per country. We observe that the majority of countries with high ranking have actually high number of locally hosted third-parties. However, Turkey, Hungary, Russia and China constitute four exceptions with over 100 locally hosted services while they are ranked below 50. Furthermore, we calculated the Pearson correlation between the number of locally hosted third-parties and the Web Index ranking. The resulting correlation is 0.26 which indicates a weak yet a positive correlation between the two factors. However, the correlation is not statistically significant (P-value=0.2009, Pearson correlation).

How are third-party trackers distributed geographically when compared with the location of the websites they're embedded on? The heat map in Figure 4.5 displays the local and overseas presence of third-party trackers. The *y*-axis shows the hosting country of the visited website and the *x*-axis shows the country of the third-party tracker. For example, the top row shows that the local websites of Qatar have embedded third-party trackers mostly from US in addition to a few from Japan. We find large presence of the overseas third-party trackers (40%=2,988 cases) across countries while some countries are clearly dominant. We identify United States, Japan, Great Britain, France and Germany as the countries with the third-parties across local websites of almost all countries in our dataset. The presence of third-parties based in the US is by far stronger than the other studied countries. This is in line with the findings reported in [8]. Amongst European countries, Great Britain, France and Germany have similar and notable presence of third-parties in each other's local websites. The same holds for the third-parties from Norway and Sweden. In the Middle East, Turkey and Israel have considerable presence of overseas third-party trackers from Netherlands, Sweden and particularly Russia in addition to the aforementioned dominant countries of Europe. In East Asia, some countries such as Japan

[2] `http://thewebindex.org/`

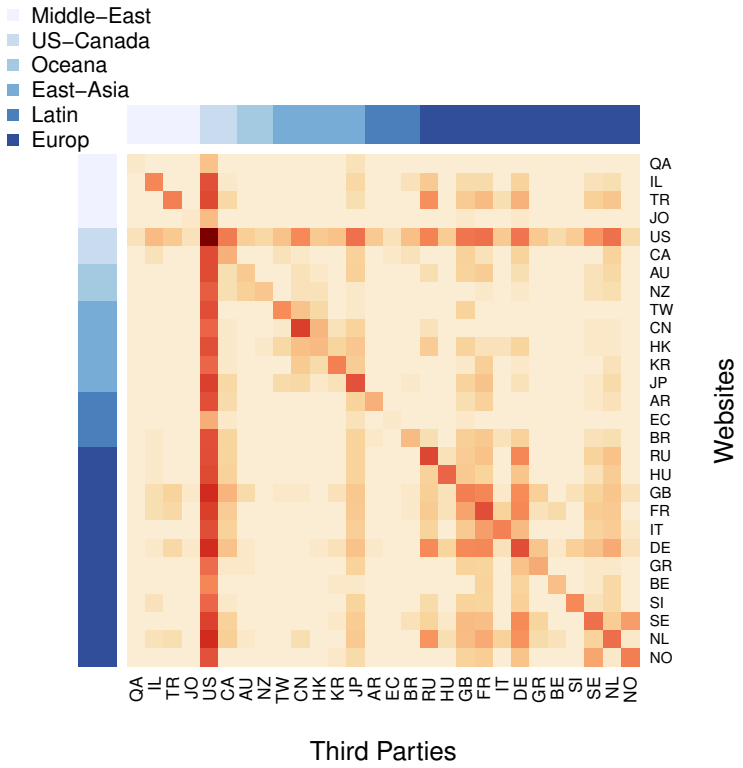


Figure 4.5: Heatmap showing locations of third-parties. Darker colours indicate greater presence, and the region of each country in the two left-most plots is depicted by the colour of the blue bars on the left and at the top.

and Hong Kong have notable presence of European third-party trackers whereas China has only a few European third-party trackers that are from Sweden (1), Netherlands (1) and Russia (2).

What is the role of third-party trackers in terms of services they provide? We categorise third-party trackers based on their provided services (see Chapter §2) using information available on Ghostery [73] and Abine[86] about tracking companies in addition to manual inspection of the third-party trackers' websites. In total we categorise 424 third-party trackers including the top 200 global and top 200 overseas ones.

We find 40% (=158) of the categorised third-party trackers providing services related to the advertisement entities (see Section §2.1.1) such as targeted advertisers, advertisement servers (hosting advertisement content), advertising exchanges (Table 4.2). The second most services are analytics (see Section §2.1.2) and Web hosting services e.g., CDNs (see Section§2.1.4) which are provided by 20% (=100) of the third-party trackers. We observe that the diversity of services provided by the local third-party trackers varies across countries. In US, Russia and France the variety of services are high, whereas in Germany, Great Britain and Japan third-party trackers are

Service	# Third-Parties
Targeted Advertisers	46
Advertising Exchanges	23
Advertisement Servers	12
Total	158

Table 4.2: Role and number of categorised third-party trackers which are part of advertisement entities

Country	Service(# websites)
JP	Advertisement Entities(526)
CN	Analytics(154), Portal(70), Shopping(57)
RU	Advertisement Entities(459), Search Services(349), Analytics(93), Hosting Services(90)
DE	Advertisement Entities(548)
GB	Advertisement Entities(326)
FR	Advertisement Entities(721), Analytics(161), Hosting Services(57)
HU	Advertisement Entities(105), News(96)
NL	Hosting Services(54), Advertisement Entities(50)
US	Advertisement Entities(28,979), Analytics(12,963), Hosting Services(3,883), Application Providers(730), OSN(238), Search Services(136)

Table 4.3: Top services provided by local third-party trackers of dominant countries.

mainly part of advertisement entities. Hungary has the highest number of news related services while portals and shopping services are considerably higher in China. We summarise the key services provided by the local third-party trackers of dominant countries in Table 4.3.

4.4 Regional Analysis

We carry on our analysis by identifying dominant third-party trackers in each region after aggregating third-parties within their parent companies, identified through a combination of three methods. First, we use Collusion's dataset [87] to detect third-parties belonging to the same company. We manually inspect this dataset for any changes using websites and Wiki pages of the involved companies. Second, we use the e-mail addresses of third-party domains obtained by querying their ADNS record (see Chapter §2). However, the email address is unhelpful if it is a general account from a cloud, CDN or DNS service. For example, `awsdns-hostmaster@`

Co.(#Domains)	**Co.(#Domains)**
Google* (42)	Amazon* (3)
Aol (18)	Facebook (3)
Yahoo* (14)	RadiumOne (3)
Sina (12)	Sizmek* (3)
Conversant (11)	AudienceScience (2)
Baidu (7)	Burst Media (2)
247 Real Media (4)	Nielson (2)
ComScor* (4)	Twitter (2)
Adobe* (3)	Quantcast* (2)
AddThis (3)	

Table 4.4: Top-20 ad related companies and number of their third-party tracker domains. * indicates companies whose trackers appeared in all countries.

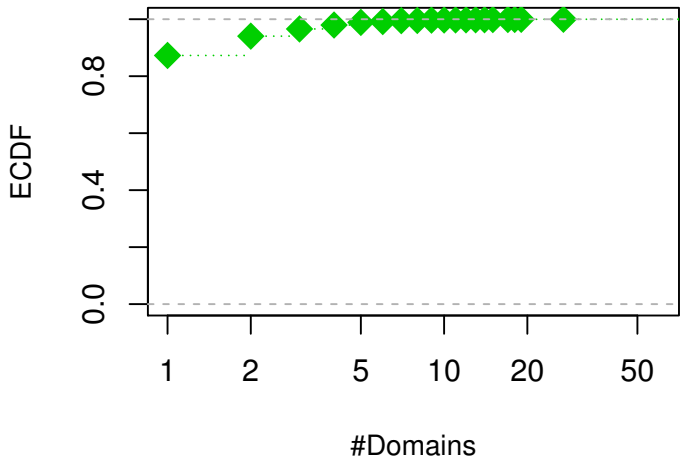


Figure 4.6: Aggregated third-party trackers within their parent companies

`amazon.com` is the email address of all third-parties hosted on Amazon Web Services, and `dns-admin@google.com` is assigned for all services hosted on Google App Engine. We identify the unhelpful email addresses by their email domain name belonging to the known CDN and DNS services, or containing keywords indicating such services. For these cases we use the organisation that is indicated in their `whois` records if available, or else we assume the third-party has no parent company. We are aware that there can be some cases with an outdated `whois` record or email addresses but we believe this is the best approach that can be executed automatically.

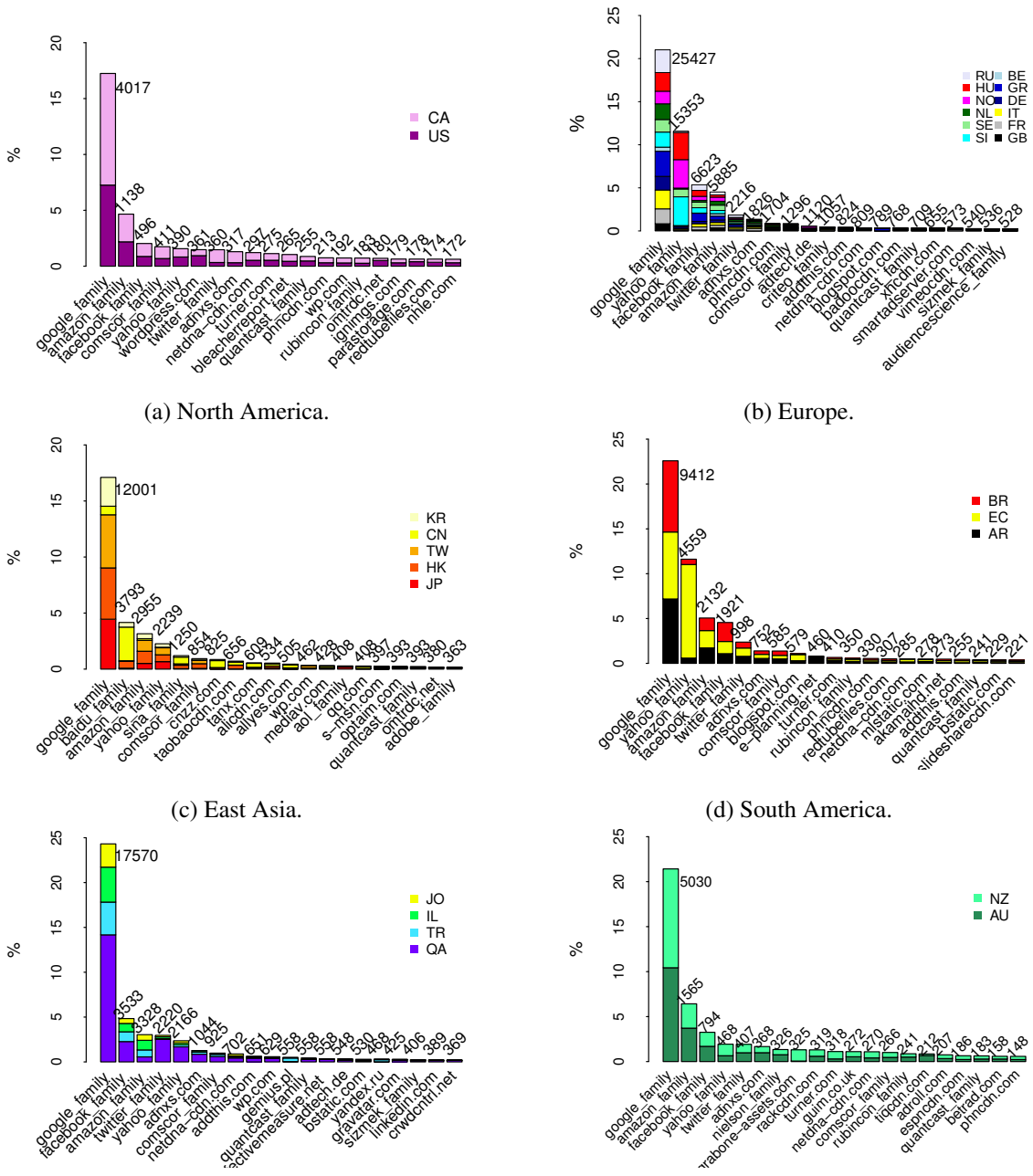
The distribution of aggregations we carry out is shown in Figure 4.6. The size of the parent companies varies considerably: some appear to own tens of third-party trackers while others have fewer than five. Table 4.4 shows the well-known companies related to advertisement entities (see

Section §2.1.1) and number of their third-party domains. We find that Google, AOL and Yahoo own the largest number of third-party trackers. The number of third-party tracking services owned by these companies has increased compared with the figures reported by [6, 2]. The third-party trackers belonging to the aforementioned companies as well as Adobe, Amazon, ComScor, Quantcast and Sizmek appears in all countries of our study. These companies are indicated by * in Table 4.4.

What are the dominant players across the geographic regions? In addition to the well-known services such as Google, we also observe some less well-known third-party services spread across almost all regions. We present the top-20 in each region in Figure 4.7. We find third-party services belonging to Google, Amazon and Facebook roughly in the same position throughout our investigated regions (top four) while Yahoo, compared with the other regions has a notably higher position (second place) in Europe (Figure 4.7b) and South America (Figure 4.7d). This difference in South America is due to the high number of occurrences of the Yahoo third-party requests in Ecuador (4,304; 10% of the third-party websites in South America). Similarly, Slovenia, Norway and Hungary contribute most in Europe (Figure 4.7b).

Beside the most famous players, we identify other third-party services with extensive presence across all regions. For example, `scorecardresearch` belongs to comScore Inc., an analytics company, `netdna-cdn` belongs to NetDNA, a CDN company, and `quantserve` belongs to QuantCast, a behavioural advertising company. These appear in almost all regions except that `netdna-cdn` doesn't appear in East Asia (Figure 4.7c). This presence implies a growing competitiveness of such businesses across the regions.

Which geographic regions have higher number of local third-parties? We observe a notable presence of local third-parties in specific regions such as East Asia and Europe. We remind that local third-parties are those services which are physically located in the related region. In East Asia (Figure 4.7c), 11 cases from the top-20 are based in this region (e.g., `sina-family`, `tabaocdn.com`); in Europe (Figure 4.7b), 4 services amongst those presented are mainly found in European countries (DE-based: `adtech.de`; FR-based: `criteo.com`, `smartadserver.com`; GB-based: `badoocdn.com`). On the other hand, in Oceania (Figure 4.7f) and South America (Figure 4.7d), there are far fewer local third-parties (one out of top-20) and in the Middle East (Figure 4.7e) there are none in the top-20.



(a) North America.

(b) Europe.

(c) East Asia.

(d) South America.

(e) Middle East.

(f) Oceania.

Figure 4.7: Top-20 third-party websites by region. Occurrence count for each third-party is displayed above each bar.

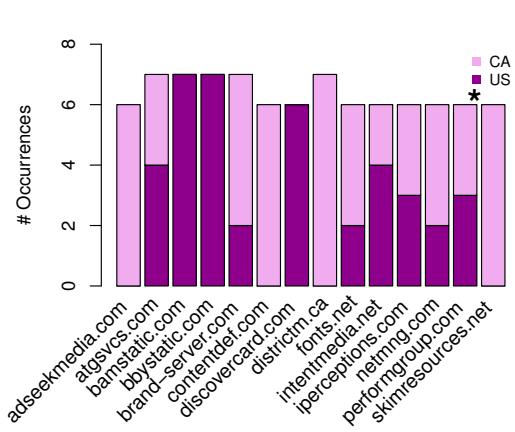
What does the ecosystem of third-parties look like if we put the dominant and popular players aside? Considering the dominance of US third-party trackers, we exclude all the local third-party trackers of the US. This leaves us with about 70% (4,505) of the total identified third-party trackers.

Figure 4.8 shows the last-15 minor third-parties in each region. While minor services have expectedly low occurrence in Europe (Figure 4.8b), North America (Figure 4.8a), the Middle East (Figure 4.8e) and Oceania (Figure 4.8f), their presence in South America (Figure 4.8d) is relatively high. Moreover, these services are equally spread across countries of each region with the exception of Qatar which has a high occurrence of the minor third-party trackers in comparison with other countries in the Middle East (Figure 4.8e).

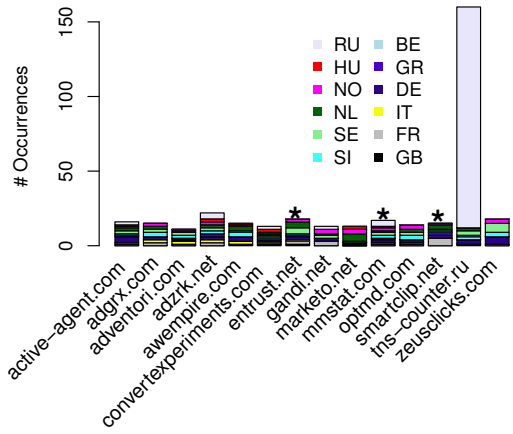
It is notable that amongst these minor services some are globally active. We identify 5 cases amongst last 15: `entrust.net` (Canada-based identity protection company), `mmstat.com` (China-based analytics), `performgroup.com` (GB-based sport content distribution and ad broker company), `smartclip.net` (German-based ad re-targeting) and `ctnsnet.com` (GB-based ad re-targeting). We don't observe any global third-party tracker in South America (Figure 4.8d) however we identify number of local and Spanish minor trackers such as `buscape.com.br` (Brazil-based marketplace) and `epimg.net` (Spain-based popular news website).

4.5 Conclusions

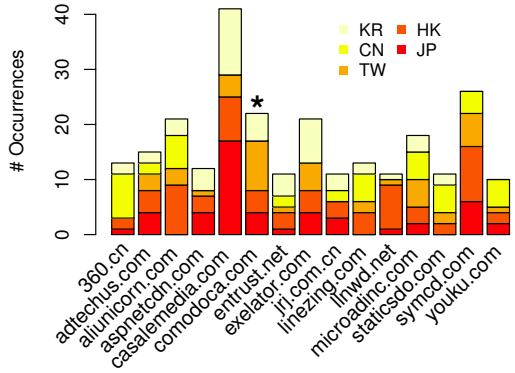
In this chapter, we presented a study of the geographic differences in the 2presence of third-party trackers. We sampled the Alexa top-500 most popular websites in each of 28 countries across widely spread regions of the world: North America, South America, Europe, East Asia, Middle East and Oceania. In line with our prior expectations (see §4.1), we identified considerable presence of local third-parties across popular websites of the studied countries, in particular in East-Asia and Europe. We identified a positive (yet small) correlation between the level of Web penetration in the countries of our study and the number of local third-party trackers of those countries. However, this correlation is not statistically significant. We identified countries such as China, Russia, Hungary and Turkey with low Web penetration ranking but considerable number of local third-parties. We think that factors such as strong state-controlled cyber security strategies can weaken this correlation.



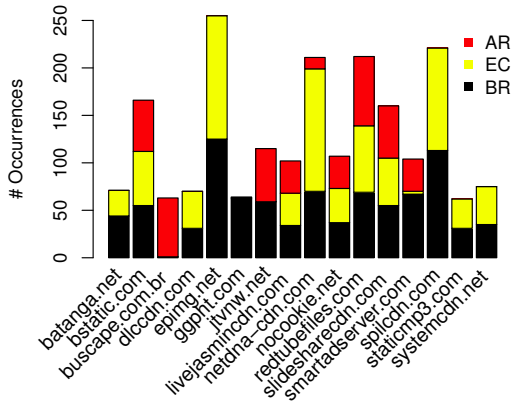
(a) North America.



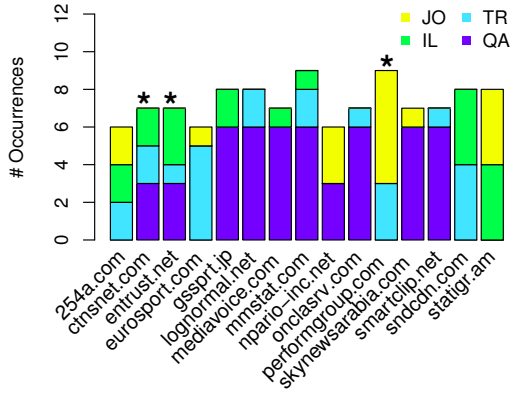
(b) Europe.



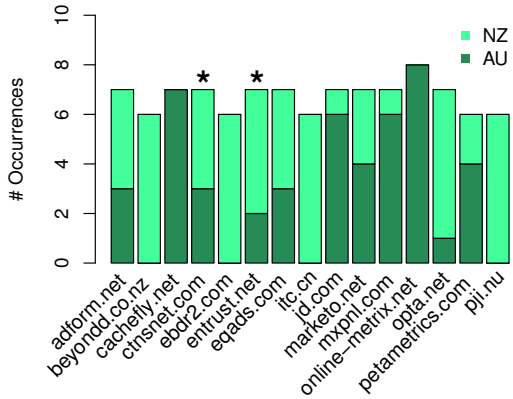
(c) East Asia.



(d) South America.



(e) Middle East.



(f) Oceania.

Figure 4.8: last-15 minor third-party trackers per region. Globally observed third-parties are indicated by *.

We have exposed the key services provided by local third-party trackers of different countries. While most of the countries have a high number of services related to the targeted advertising, some appear to have a high variety of services. One example is Russia with its high presence of advertisement entities, search and analytics services. Having various type of services spread across popular websites of different countries enables one country to access different sorts of user information belonging to citizens of other countries. For instance, in case of Russia its local third-party trackers have considerable presence in popular websites of Germany, Netherlands, US and Turkey. This presence implies the potential access of Russia to the data of citizens of the aforementioned countries which makes legal and financial management of personal data flow challenging. In Chapter §7 we discuss some of these challenges.

Chapter 5

Tracking Personal Identifiers Across the Web [1]

5.1 Introduction

In the previous chapter, we showed a strong presence of local and global third-party trackers across popular websites of various countries and regions. In this chapter, we focus on the third-party trackers assigning personal identifiers to users and share these identifiers with other parties.

One of the techniques used by third-party trackers to follow user's activities across the Web relies on sharing *user-specific* identifiers (IDs). Figure 5.1 shows an example of this practice happening when a user visits a website such as `cnn.com`. In this scenario `rubiconproject.com`, a third-party tracker for `cnn.com`, shares its user-specific identifier with another third-party tracker such as `adrate.com` via the `rub-Id` URL parameter. In this example, `rubiconproject.com` and `adrate.com` are able to merge their datasets based on the shared ID.

As we reviewed in §3.3, a few works highlight the presence of user-ID sharing [6, 17]. The presence of this practice particularly between advertisement related parties is reported in [17]. Moreover, the authors in [38] introduce a method to identify user-specific IDs (see §3.3). They identify 730 parties that are involved in sharing user-specific IDs. Considering the aforementioned findings, we expect that this practice is dominant across various players of different com-

[1] This study has been published in the proceedings of the 12th Conference on Passive and Active Measurement (PAM),2016, Cyprus [81]

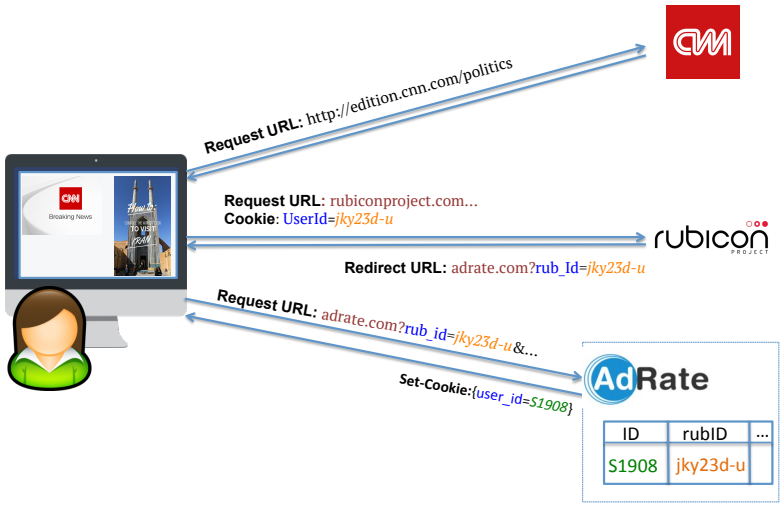


Figure 5.1: An example in which `rubiconproject.com` shares its user-specific identifier with `adrate.com` while a user is visiting `cnn.com`

.

panies involved with targeted advertisement services. Furthermore, We expect that factors such as profile size (e.g., amount of their browsing history) and profile condition (logged-in or logged-out) affect the presence of ID-sharing.

In order to evaluate our expectations, we explore the characteristics of user ID-sharing groups by analysing the organisational and categorical relation amongst the members of ID-sharing groups (§5.2). Afterwards in Section §5.3, we investigate the effect of user profile on the presence of ID-sharing groups.

5.2 User Tracking

We start our analysis by exploring the connections between domains when they are aimed to track users. User tracking is a practice by which a domain, either being directly visited by a user or indirectly through third-party trackers, assigns a unique identifier to the user, and shares this identifier with other domains. The parties participating in user tracking are able to aggregate the data collected by other parties in order to construct a comprehensive profile of users. In the rest of this section, we first describe our methodology and dataset, and subsequently explore the size and nature of a user ID-sharing group.

5.2.1 Methodology and Data Collection

Region	Country	No. P
Europe	BE, CH, FE, DE, DK, ES, GB, GR, NL, FI	97
Middle East	BD, IR, QA	11
America	CA, MX, US	9
Oceania-East Asia	AU, CN, MY	8
Africa	SG, MR	3

Table 5.1: Number of participants per geographical location.

URL	User-Specific IDs	Key
http://ads.rubiconproject.com/ad/11078.js	65d39451-1f73-435a-bf39	put_2760
http://apex.go.sonobi.com/trinity.js	i736hcjtwb05natk	uin_bw
http://cm.adform.net/pixel	d4848—VOzy0—N1xas	adform_pc

Table 5.2: Example of URLs and the identified user-specific IDs with their associated keys.

In order to collect data, we use the same Firefox plug-in that we used before in Chapter §4. Our plug-in records all HTTP requests and responses passing through the browser. The functionality of our plug-in is described in Section §4.1.We ask our participants to install our plug-in and use Firefox as their main browser for the minimum duration of two weeks. In order to preserve users' privacy we do not record any identifiable information such as the IP address, name or contact information. Additionally, we obtained ethics approval from QMUL ethics committee (code QMREC1416a) before performing our user studies. All our data are obtained between 20 February 2015 until 1 April 2015. In total we have 129 participants from 22 countries across the globe. Our participants have visited 4,951 unique websites which include 6,568 unique third-party trackers. Table 5.1 lists the number of our participants in each geographical region.

5.2.2 Nature of ID-Sharing Groups

To explore user tracking via sharing user-specific identifiers, we first need to determine the identifiers that are likely to be used as *user-specific IDs*: a unique identifier stored in a cookie or embedded as a parameter in a URL. For this purpose, we apply the following rules inspired by Acar *et al.* [38] on all items stored in the cookies and the URL parameters.

- Extract `(key,value)` pairs using delimiters such as ampersand (`&`) and semi-colon (`;`). For instance, this string `id=ece53b2e-ea5c-4433-ad3d&ssid=02ba238451c-ec44ba88` contains two (key,value) pairs: (id,ece53b2e-ea5c-4433- ad3d) and (ssid,02ba-

238451cec44ba88).

- Exclude (key,value) pairs that are *inconsistent*: a (key,value) pair is inconsistent if there are multiple values for the same key belonging to a certain domain. For example these pairs (id,ece53b2e-ea5c-4433) and (id,ffc87j3o-gh11-3278) observed from `bbc.co.uk` are excluded.

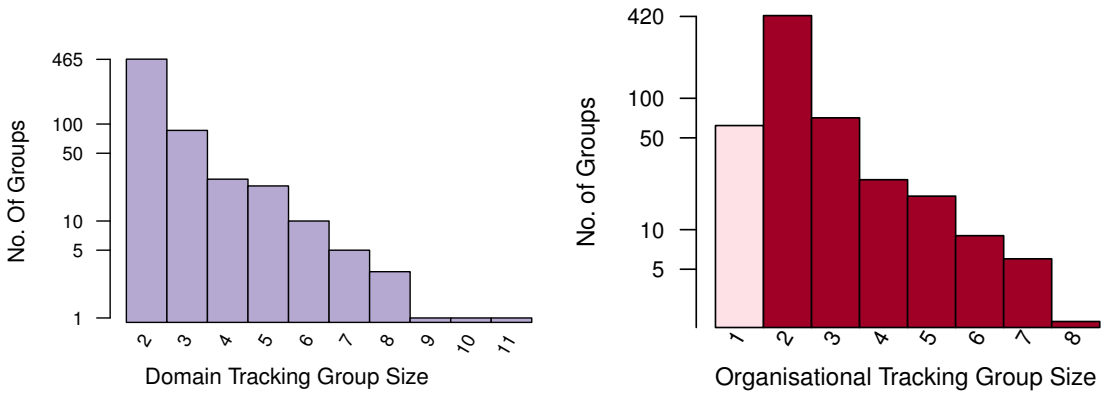
- Exclude those value strings that are shared by multiple users.

- Only include those value strings that their length is longer than 7 characters. After applying the aforementioned rules on our dataset, we found that 96% of user-specific IDs have a minimum length of 7 characters.

We apply the above-described method for each user. Table 5.2 shows sample URLs and their identified user-specific IDs with their associated keys. The identified IDs appear in various formats of which the most common are {`xx..x`}, {`x-x-..-x`} and {`x|x|..|x`} where `x` can be any combination of characters and numbers. We find 3,224 unique user IDs from 806 domains. The vast majority of these IDs (96%) are being shared between at least two domains. We identify 769 domains that share unique user IDs with other domains. Extracting the user-specific IDs enables us to identify *user ID-sharing groups*: a set of domains that share user-specific IDs. We identify 660 unique ID-sharing groups containing two to more than eight domains. Figure 5.2a provides the distribution of the number of different sharing groups (y-axis uses a logarithmic scale) across their group size (x-axis). From Figure 5.2a, we observe that user IDs are mainly shared between two (467 unique groups, 2,742 occurrences) or three (86 unique groups, 201 occurrences) domains. Moreover, the number of unique groups and their occurrences drop steadily as group size increases.

5.2.2.1 Organisational Sharing

User ID-sharing groups consist of multiple domains that may actually belong to the same organisation. Therefore, we broaden our approach from domains to organisations, resulting in *organisational sharing groups*. For example, the organisational sharing group for {`google.com`, `youtube.com`} is {Google}, and for this group: {`youtube.com`, `scorecardresearch.com`} is {Google, comScore}.



(a) Domain ID sharing groups

(b) Organisational ID sharing groups

Figure 5.2: Size of ID sharing groups based on number of (a) domains and (b) organisations (the highlighted bar shows within organisational sharing). Y-axis in both figures uses a logarithmic scale.

We identify the organisation behind a set of domains using the method described in the Section 4.4. Figure 5.2b provides the distribution of the number of organisational sharing groups (again using a logarithmic y-axis) across their sizes (x-axis). The highlighted bar shows within-organisational sharing groups.We observe that the number of within-organisational sharing groups (sharing within a single organisation) is considerably lower than those with more than one organisation (sharing across different organisations). Moreover, the most cross-organisational sharing appears between only two organisations. The majority of these two- organisation groups contain a member organisation that appears only once (306). On the other hand, dominant organisations such as Google, Rubicon Project and Optimizely (a user targeting company) appear in 43, 40 and 33 two-organisation groups respectively.

In general, we find some organisations such as Rubicon Project (an ad exchange company) appears strongly in the cross-organisational sharing groups (112 groups) while large organisations such as Google appears in both cross-organisational and within-organisational sharing groups. Table 5.3 shows the top-15 most popular organisational sharing groups (in their frequency of occurrence) and the nature of their user-specific ID-sharing within the group, i.e., within an organisation (w-org) or cross organisations (c-org).

5.2.2.2 *Cross Categories Sharing*

To gain more insight into the nature of user ID-sharing, we analyse the ID-sharing groups with a different approach. We examine the categories of domains in each group. We first identified

Sharing Group	Type
google.com, googleadservices.com	w-org
google.com, youtube.com	w-org
flickr.com, yahoo.com, yahooapis.com	w-org
bbc.com, effectivemeasure.net	c-org
yahoo.com, yimg.com	w-org
bing.com, live.com	w-org
adxcore.com,cherryssp.net	c-org
rubiconproject.com, wtp101.com	c-org
rubiconproject.com, tapad.com	c-org
bing.com, live.com, msn.com	w-org
eyeviewads.com, rubiconproject.com	c-org
everesttech.net, rubiconproject.com	c-org
rubiconproject.com, w55c.net	c-org
sina.com.cn, weibo.com	w-org
rubiconproject.com, rundsp.com	c-org

Table 5.3: Top 15 user ID-sharing groups ordered based on their frequency of occurrence. The Type column indicates the nature of organisational sharing within the group (within-organisation=w-org versus cross-organisation=c-org).

domain categories using the Trend Micro Site Safety Center categorisation service[88]. The Trend Micro service contains 85 different interest categories. Moreover, we manually inspect those that were not available on Trend Micro. We find categories related to the advertisements (e.g., advertising networks, analytics, advertising exchanges) have the highest presence. This strong presence is expected due to the employed advertising mechanisms (e.g., real-time bidding) that share user-specific IDs across different entities of the advertisement related services.

We then compare the categories of domains in each group. For instance, in the following ID-sharing group {`getclicky.com`,`ibtimes.co.uk`} the categories of domains in the group are {`Analytics`, `News`}. Table 5.4 shows the top 15 categories of the sharing groups (in their frequency of occurrence) and the nature of their domain categories in the group i.e., within a category (w-cat.) or cross categories (c-cat). We observe that the majority of ID-sharing in the groups happens across different categories. We find only 28 ID-sharing groups of which their members belong to the same category (within-category sharing). This number is considerably lower than 110 groups with members belonging to different categories (cross-categories sharing). We also observe that sensitive domain categories such as health related ones participate in the ID-sharing with domains related to advertisement trackers and search engines (7 groups). For instance, `webmd.com` (a health information website) has shared user-specific IDs with `gravity.com` (an advertisement tracker). Looking at a sample HTTP request from

Sharing Group	Type
search engines, web advertisements	c-cat.
search engines, streaming media	c-cat.
ad-tracker	w-cat.
search engines	w-cat.
ad-tracker, web advertisements	c-cat.
ad-tracker, internet infrastructure	c-cat.
ad tracker, photo searches, search engines	c-cat.
media, news	c-cat.
ad tracker, news	c-cat.
web advertisements	w-cat.
ad-tracker, business	c-cat.
health	w-cat.
internet infrastructure, web advertisements	c-cat.
ad tracker, search engines	c-cat.

Table 5.4: Top 15 categories of the sharing groups ordered based on their frequency of occurrence. The Type column indicates the nature of domain categories within the sharing group (within category=w-cat. versus cross category=c-cat.).

RequestURL:	`http://rma-api.gravity.com/v1/beacons/log?action=beacon&`*user_guid*=**21737bfabd4416779f6**&*referrer*= **http://www.webmd.com/search/search_results/default.aspx?query=breast-cancer**
Host:	`rma-api.gravity.com`
Referrer:	`http://www.webmd.com/breast-cancer/default.htm`

Table 5.5: A sample HTTP request from webmd.com (a health information website) to gravity.com (an advertisement tracker). Gravity.com logs users' visited pages via *referrer* URL-parameter. Consequently, the searched terms by users on webmd.com are exposed to gravity.com (e.g. query=breast-cancer)

`webmd.com` to `gravity.com` in Table 5.5, shows that `gravity.com` logs users' visited pages via *referrer* URL-parameter. This information enables `gravity.com` to create users' profiles based on their visited pages and searched terms on `webmd.com`. The presence of such domain categories within sharing groups raises serious privacy concerns since users' sensitive information can be exposed within sharing groups.

5.3 Effect of User Profile

In the previous section, we observed strong presence of user ID-sharing based on two-weeks online activities' logs of over 100 users. In this section, we further examine the potential intentions behind the ID-sharing by studying the effect of user profile on the presence of ID-sharing domains. For this purpose we run multiple experiments on sets of trained user profiles. In order

Profile Size	#Domains
P-500	649
P-200	631
P-0	538

(a) Profile Size

Profile Condition	#Domains
no-account	531
logged-in	599
logged-out	749

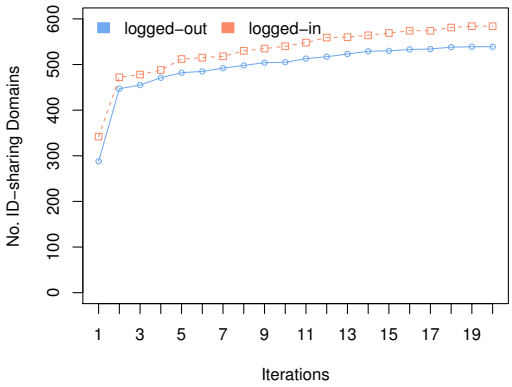
(b) Profile Condition

Table 5.6: Total number of unique ID-sharing domains for each (a) profile size and (b) profile condition.

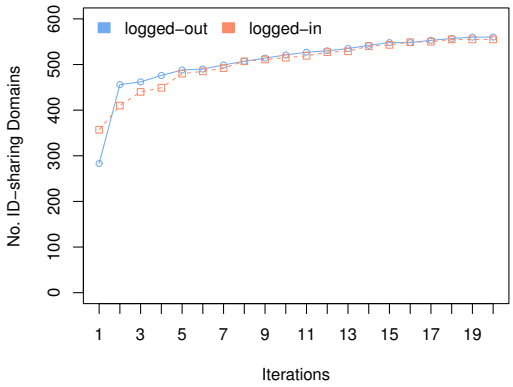
to create the user profiles, we first created five artificial users with separate accounts on Google, Amazon, eBay and Twitter. We assign three different profile sizes, in terms of the browsing histories, to our users: (1) Two users were given a browsing history consisting of Alexa's top 500 websites (*Profile-500*); (2) Two other users with smaller size of browsing history including the Alexa top-200 websites (*Profile-200*); (3) One user with an empty browsing history (*Profile-0*). To explore the effect of not having a user profile, we consider a user with an empty browsing history and without any accounts on the aforementioned websites (*noAcount*). We create the browsing history by crawling the corresponding Alexa's list of websites for five consecutive times while users were logged-in. The profile-training step is done on the Firefox browser installed on a separate Linux machine per user. After creating the user profiles, we install the Firefox plug-in from the section 5.2.1 on the Firefox browsers. Then, we execute the main step of the experiment by visiting Alexa's top 1000 websites for each user. We repeat this step for 20 iterations to expose as many as possible ID-sharing domains. We perform the main step identically under two conditions: user logged-in and user logged-out.

We apply the same rules as described in Section 5.2.2 to identify user-specific IDs. Consequently, we identify 4,104 unique user-specific IDs shared by 787 domains. Figure 5.3 illustrates the accumulated number of unique ID-sharing domains across the iterations per user and profile condition. We observe that the highest rise occurs between the first and second iteration (approximately 40%), in comparison with subsequent iterations (Figure 5.3).

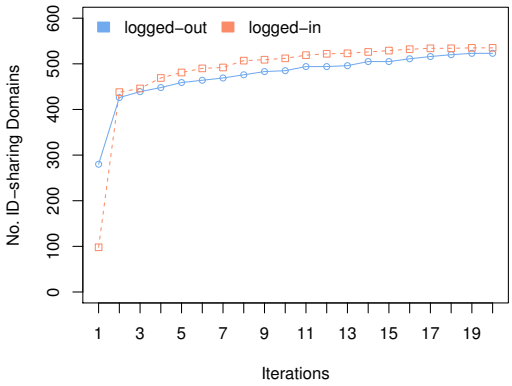
Moreover, we explore the number of ID-sharing domains across various profile sizes (browsing histories) and profile conditions (logged-in, logged-out, and noAcount). Table 5.6 shows the unique number of ID-sharing domains per profile size and condition. The results in Table 5.6 suggests that users with a larger profile (more browsing history) are tracked by a higher number of ID-sharing domains than those with smaller profile sizes. This result was expected since larger



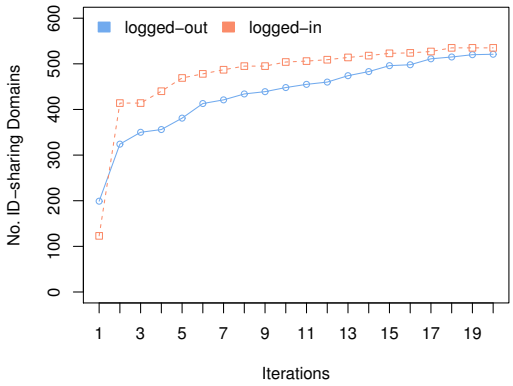
(a) Profile Size: 500



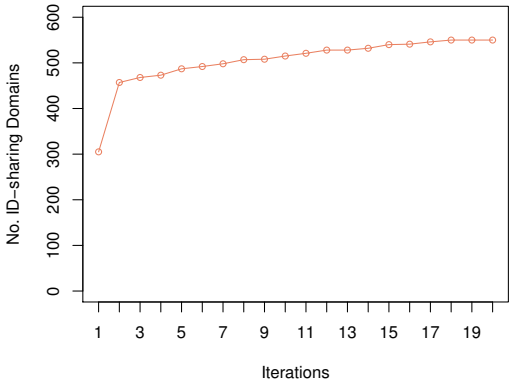
(b) Profile Size: 500



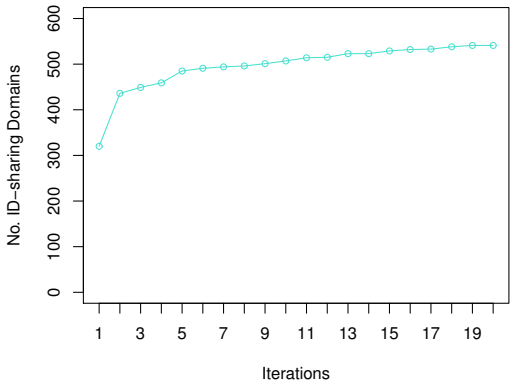
(c) Profile Size: 200



(d) Profile Size: 200

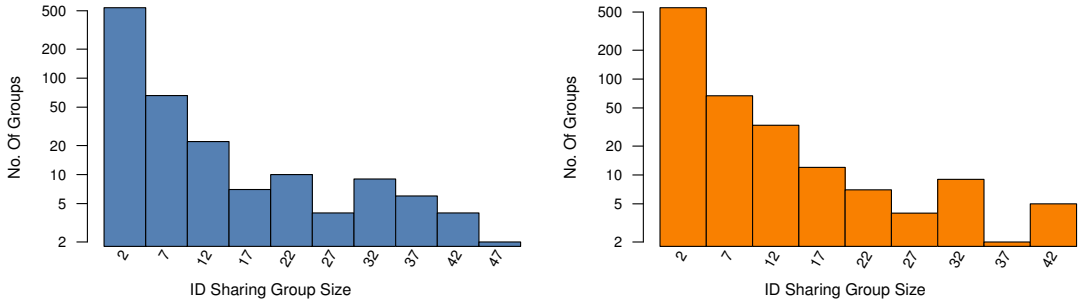


(e) Profile Size: Empty



(f) Profile Size: Empty and without any account)

Figure 5.3: Number of ID-sharing domains across the iterations for different profile sizes and profile conditions (logged-in vs. logged-out)



(a) Logged-out (b) Logged-in

Figure 5.4: Organisational ID-sharing groups across various profile conditions: (a) logged-out and (b) logged-in (Y-axis in both figures uses a logarithmic scale).

profiles expose more information to third-party trackers. On the other hand, we find the number of ID-sharing domains higher in the logged-out condition than logged-in (Table 5.6b). In general, the comparable numbers of ID-sharing domains across various profile conditions and profile sizes suggest that the users are being tracked regardless of their profile condition and the amount of the browsing history (Table 5.6).

Afterwards, we examine the presence of organisational ID-sharing groups across different profile conditions. We define ID-sharing groups as sets of domains that share user-specific IDs (refer to Section 5.2.2). In addition, we identify the organisations behind the sharing groups using the method described in the Section 4.4. We identify 694 ID-sharing groups of which 357 (=51%) belonging to two distinct organisations. We find that across these groups, Google and Rubicon Project have the highest presence with respectively 27 (=7%), 20 (=5%) cases. Figure 5.4 shows the number of organisational ID-sharing groups against their group size when the user is logged-out (Figure 5.4a) and logged-in (Figure 5.4b). The number of ID-sharing groups with a larger size are higher in the logged-out condition comparing to the logged-in condition. As an example, Figure 5.5 shows the largest ID-sharing group for the logged-out mode. In this group, we find the Rubicon Project, Switch Concept (an ad. Network company) and StickyADStv (a video publisher company) as the most dominant ones in terms of organisational ID-sharing. We observe strong collaborations between specific organisations such as the Rubicon Project, Sovrn (an ad Network company), Google and StickyADStv.

This finding can be due to the fact that more domains have been collaborating with each other when the user was logged-out, to compensate for the lack of context about the user, and trying to create a more precise profile for that user—by gathering as much information as possible.

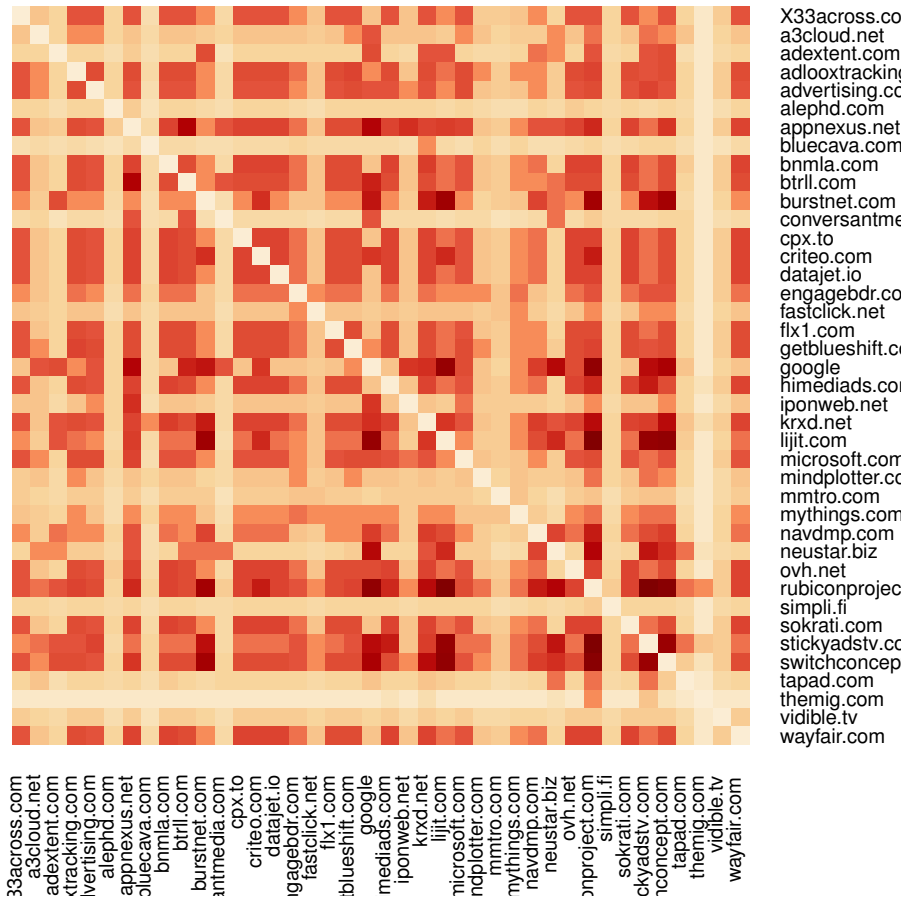


Figure 5.5: Heatmap showing the biggest organisational ID-sharing group in the logged-out mode. Darker colours indicate higher frequency of collaboration between two organisations.

5.4 Conclusion

In this chapter, we explored the entangled connections between all user tracking parties on the Web. In particular, we investigated the tracking groups that shared user specific identifiers. We recorded the browsing history of more than 100 users for more than two weeks. We find 660 ID-sharing groups in our data. We identify a significant amount of ID-sharing across different organisations. The number of ID-sharing within organisations are considerably lower. We note that within-organisational ID-sharing can happen through other networks to which we did not have access (e.g., company intranet). We identified Google and Rubicon Project (an ad. network company) as the most dominant companies that used ID-sharing. Similar to our observation at the organisational level, we observe a significant presence of domains from different categories within ID-sharing groups. We observe that sensitive domain categories such as health related ones participate in the ID-sharing with domains related to advertisement trackers and search engines (seven ID-sharing groups).

Moreover, we examined the effect of user profile on the presence of ID-sharing domains. Contrary to our initial expectation (see §5.1), the changes in the number of ID-sharing parties across various profile sizes are comparable. This suggests that users are being tracked regardless of the amount of their browsing history. We observe that the number of ID-sharing domains are higher in the logged-out condition than logged-in. Our results suggest that more domains are collaborating with each other when the user is logged-out trying to create a more precise profile for that user.

This work can be extended by investigating whether this collaboration amongst ID-sharing domains in the logged-out mode aims to identify the user, or it is a side-effect of knowing less about the user, hence being more inclusive in potential advertising sources. Note that from our data we cannot directly observe whether domains use these IDs to merge collected data from different sources. However, considering the possibility of such practice, we believe it is important to get additional insight about what ID-sharing groups actually do through the user IDs.

Chapter 6

The Effect of Blocking Trackers on Page-Load Performance [1]

6.1 Introduction

In the previous chapter, we showed that users are being tracked regardless of the amount of their browsing history. In response, users can turn to tracker-blockers to preserve their privacy and improve their browsing experience. In this chapter, we focus on the effect of tracker-blockers on page-load performance of websites.

A recent study has estimated the number of active tracker-blocking users to be 198 million [75]. However, some websites depend on third-party trackers for delivering their services. Such a collaboration can be endangered in the presence of tracker-blockers and may lead to a reduced website functionality and performance. For example, consider two JavaScript resources of a website as shown in Figure 6.1. `ad.js` is in charge of delivering advertisements while `loader.js` is in charge of loading all images of the website (including the advertisement images). In this scenario, `loader.js` is partially dependent on the output images of `ad.js`. If an tracker-blocker blocks `ad.js`, `loader.js` will wait for the output images of `ad.js` until it times out. These dependencies amongst different resources, in particular JavaScript resources, play a significant

[1] In this chapter the findings of [89] is used to analyse the performance of websites that differentiate between users with and without tracker-blocker.

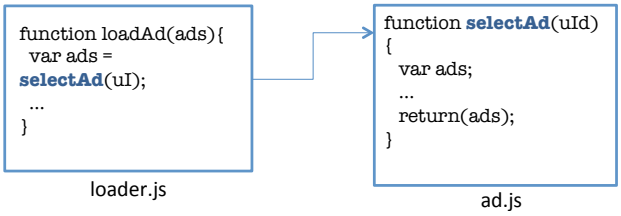


Figure 6.1: Dependency between JavaScript resources of a Web page.

role on websites load time [90].

There are multiple reports emphasising the importance of websites performance on users' Web browsing experience. Amazon has reported that an increase of 100ms in page-load time can lead to 1% loss in sales. Moreover, Google stated that increasing page-load time by 500ms leads to 25% drop in number of searches [91, 92]. There are various attempts to boost websites performance by identifying the bottlenecks on the browser side, suggesting pre-processing proxy engine and new frameworks to load resources [90, 93, 94].

We expect that blocking some third-party services by tracker-blockers affects the functionality of some websites (see section §6.2), and in turn reduces their performance. However, it is expected that popular websites take into account the scenarios in which their visitors use tracker-blockers. These websites can consequently adapt existing solutions to avoid or reduce the negative effect on page-load performance due to blocking third-party services. Otherwise the reduced performance can lead to dissatisfaction of some visitors and potential financial loss for the websites.

In this chapter, we quantify the impact of tracker-blockers on the page-load time. We study the effect of two popular tracker-blockers, Ghostery and AdBlock Plus, on the performance of Alexa's top-200 websites to:

- understand whether different tracker-blockers have similar effect on the page-load performance.
- investigate the impact of tracker-blockers on popular websites.
- understand which categories of websites are affected the most.

In the following sections, we first provide background information on Web page-load process and tracker-blockers functionality (§6.2), afterwards we describe our data collection method (§6.3).

We then present our analysis on the changes of page-load performance in the presence of Ghostery and AdBlock Plus (§6.4). We show a considerable negative effect of tracker-blockers across various websites regardless of their popularity and category of services.

6.2 Background on Page Loading and Tracker-blocking

When a user enters the URL of a website into a browser, a request to download the main HTML structure of the website will be sent to the website's corresponding server. The HTML structure is dynamically generated (or it has a predefined static structure) at the server side and will be sent back the client's browser in chunks. Upon receiving the first portion of the HTML of the page, the browser recursively parses all HTML elements and process their corresponding objects. These objects include images, videos, Cascading Style Sheets (CSSs) to format the appearance of the HTML elements, JavaScripts to interact with the page at the client side. The browser gradually transfers the processed objects into another format i.e., DOM (Figure 6.2) tree which will be used later to render the page on the user's screen.

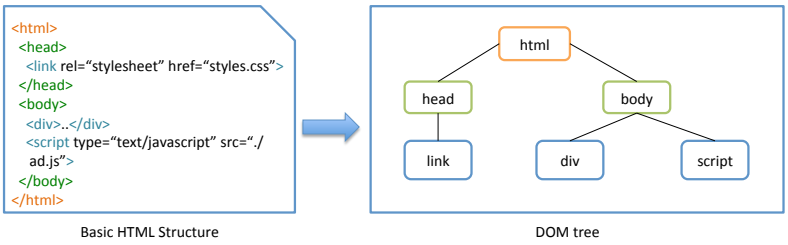


Figure 6.2: Browser engine transfers HTML structure to DOM tree.

The browser does not download and parse all of these objects in parallel. When a browser encounters the first <script> element, it will halt DOM construction until the corresponding JavaScript code of the <script> element is parsed and executed. This is due to the fact that JavaScript code can modify the HTML structure and content of the page, for instance a JavaScript code which inserts an advertisement's image into the page (Figure 6.3). The other object suspending the cycle of page load is CSS. Similar to JavaScript, CSS affects the HTML structure by modifying the format of elements (Figure 6.3). To avoid the delay caused by the parse-blocking effect of JavaScript and CSS, Web developers are advised to separate any unnecessary JavaScript and CSS from the main HTML structure of the websites [95]. Therefore, these objects will be loaded later after the DOM tree is constructed.

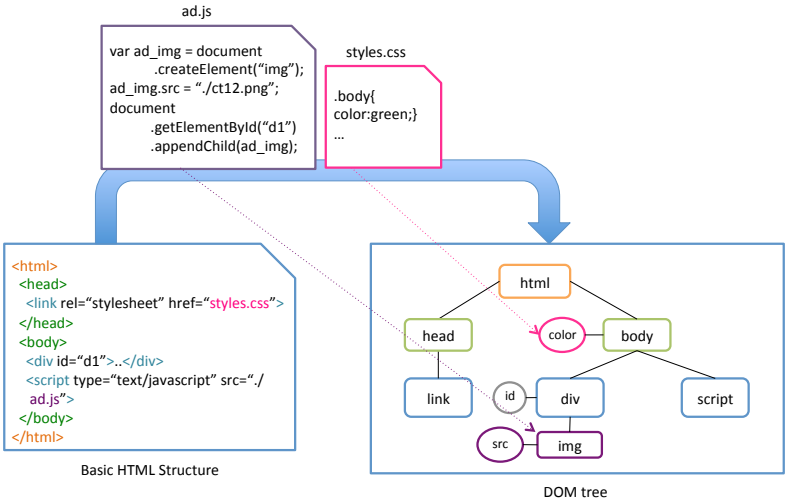


Figure 6.3: JavaScript and CSS can change the structure of DOM tree.

Tracker blocking. Tracker-blockers are aimed to raise users' awareness about third-party trackers, and improve their browsing experience by blocking tracking services. Ghostery [73] and AdBlock Plus [96] are amongst the most popular tracker-blocking plug-ins [97, 98]. The tracker-blocker browser plug-ins are able to monitor all HTTP(S) requests and responses pass through the browsers. For instance, `http-on-modify-request` is a predefined event available on Mozilla Web API through which HTTP requests can be modified or cancelled before being sent by the browser. In order to decide which requests to block, these plug-ins mainly rely on static lists of patterns referring to the user tracking resources. For example, this pattern:`http://*/ads.js` refers to the `ads.js` JavaScript code that is served by any domain. While some plug-ins such as Ghostery provide an informative categorisation of the trackers and enable users to block different categories, there are other ones like AdBlock Plus that mainly focus on blocking advertisement related trackers. In to blocking the trackers, Ghostery claims to improve page-load performance of the websites. We note that Ghostery and AdBlock Plus block trackers only if the user has enabled the blocking option and applied the corresponding configurations.

6.3 Methodology

Data collection. The data collection procedure includes two phases: first, we need to record the loading time of a website under the *standard* condition i.e., there is no tracker-blocker running on the browser. Second, we record the loading time of the same website while a tracker-blocker (Ghostery or Adblock Plus) is running on the browser. We describe these phases in the rest of

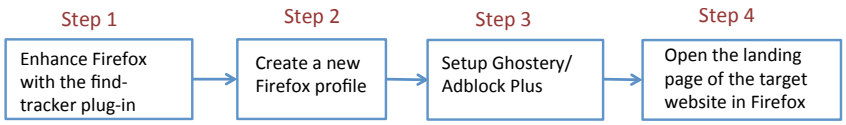


Figure 6.4: The data collection procedure

this section.

We extend the find-tracker Firefox plug-in described in Section §4.2 to record the loading time of web pages using `NavigationTimingAPI` [99]. This API provides the information about the time at which certain events happen during processing of a web page. For instance, `navigation Start` property of the Navigation Timing API represents the moment at which the browser is ready to send a request to get HTML structure of a page. Another example is `domComplete` which indicates the time at which all the resources of the Web page are parsed and transferred into DOM tree format(§6.2). We run our experiments on a Firefox browser instrumented by our plug-in. We use a similar data collection procedure as the one described in §4.2. We visit the landing page of Alexa's top-200 websites. We ensure to allocate enough time to each website to reach the `domComplete` point. Thus, we set an interval of 60 seconds for each visit. Note that the average page-load time that has been reported from different resources is less than 10 seconds ([90, 100]). If `domComplete` event of a website is not triggered within 60 seconds, we assume that there is a technical problem. We use a new browser profile for each visit to avoid the effect of caching.

We visit each website for ten times. We note that we cannot find a statistically significant difference in the average PLT time across the ten iterations (P-value = 0.9, ANOVA). Due to this observation and similar previous work done in [101], we also set the number of the iterations to ten. We report the median Page-Load Time (PLT) from a total of ten runs as this metric can represent the majority of the observed PLT times and is not influenced by extreme values (e.g., a very high PLT of a website at a certain time due to network glitches). We refer to this part of experiment as the *standard* condition (there is no tracker-blocker employed on the browser).

To measure the effect of tracker-blockers, we add a new step to our original data collection procedure (described in Section §4.2). The new step, as it is shown in Figure 6.4 (step 3), includes the installation of Ghostery and activating its blocking option for all third-party trackers (see §6.2) before visiting a website. Afterwards, we identically repeat our experiment (visiting Alexa's

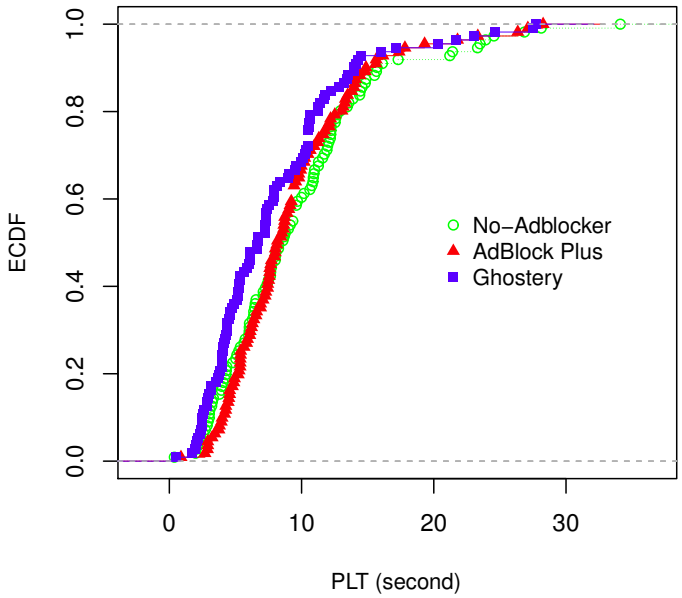


Figure 6.5: PLT comparison in the standard condition (No-Adblocker) vs. in the presence of tracker-blockers including AdBlock Plus and Ghostery.

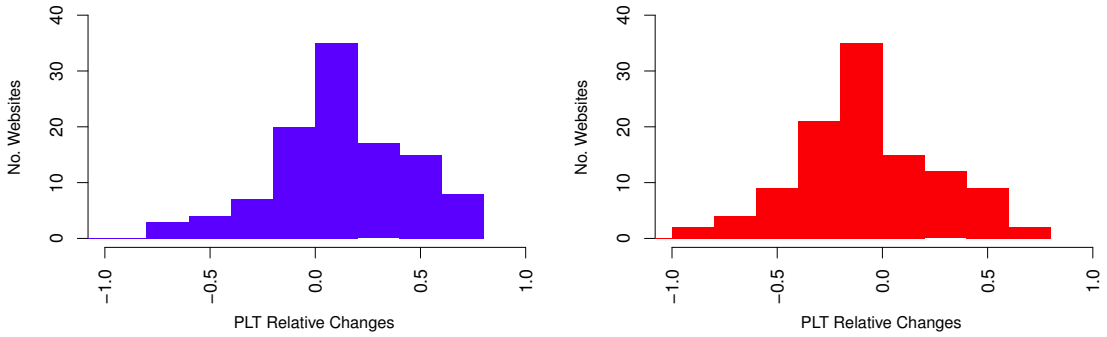
Top-200 for ten times). To compare the effect of different tracker-blockers, we similarly activate AdBlock Plus, and repeat the experiment.

We conduct our experiments on a MacBook Pro with a 2.3GHz Intel core i5 CPU and 8GB memory. The computer is connected to a home network of 20Mbps (provided by Sky Broadband service which is one of the best low cost broadband services in UK ([102]).

Page-load time. We calculate PLT as a time from which a page is requested until all page's objects are fetched, processed and added to the DOM tree (§6.2). Another approach is to define PLT only based on visual properties regardless of the interactions between browser and website. For example, *above-the-fold* time shows the time at which a website is visible on the user's screen. This metric needs to be calculated manually after the Web page is visually recorded, and thus is not scalable.

6.4 Effect of Tracker-blockers on Page-Load Performance

We aim to understand the impact of blocking third-party trackers by tracker-blockers on the page-load performance. Figure 6.5 shows the comparison of the standard PLT (no tracker-blockers) with the PLT affected by Ghostery and AdBlock Plus. We observe that AdBlock Plus and Ghostery have different effects on the page-load time in comparison with the standard condi-



(a) Ghostery (b) AdBlock Plus

Figure 6.6: Relative changes of the standard PLT caused by Ghostery (6.6a) vs. AdBlock Plus (6.6b).

tion. Ghostery has an improving effect on the performance of the websites by over 2s reduction in the total median PLT from 7.33s to 5.04s, whereas AdBlock Plus reduction of the total median PLT is 167ms. This finding is expected as Ghostery blocks various categories of trackers in comparison with AdBlock Plus which only targets advertisement related trackers (§6.2). Figure 6.6a shows the histogram of relative changes of PLT at the presence of tracker-blockers for the Alexa's top-200 websites. The relative changes of PLT is calculated using Equation 6.1 in which $PLT_{tracker\text{-}blocker}$ is the PLT of loading the website in the presence of a certain tracker-blocker.

$$\frac{PLT_{StandardCondition} - PLT_{tracker\text{-}blocker}}{PLT_{StandardCondition}} \tag{6.1}$$

Ghostery has a positive effect on the page-load performance of 103 (=64%) websites of which `forbes.com`, `dailymotion.com` and `msn.com` have the highest relative reduction of PLT with respectively 79%, 77% and 74%. Moreover, the PLT of about 60 websites (=38%) reduces by up to 50% when Ghostery is activated. Despite Ghostery's claim to optimise website performance, Ghostery exhibits a negative effect on the page-load performance of 52 websites (=32%) by increasing their PLT in comparison with the standard PLT. Comparing Ghostery and AdBlock Plus (Figure 6.6b), we find that there are only 53 websites (=33%) benefiting from AdBlock Plus of which `msn.com`, `indiatimes.com` and `aol.com` have the most relative reduction of PLT with respectively 64%, 63% and 59%. AdBlock Plus increases the PLT of 107 (=66%) websites which is two times more than the negative effect of Ghostery on the page-load time. We find 49 websites that their PLT increase by both Ghostery and AdBlock Plus of which `amazon.co.uk`

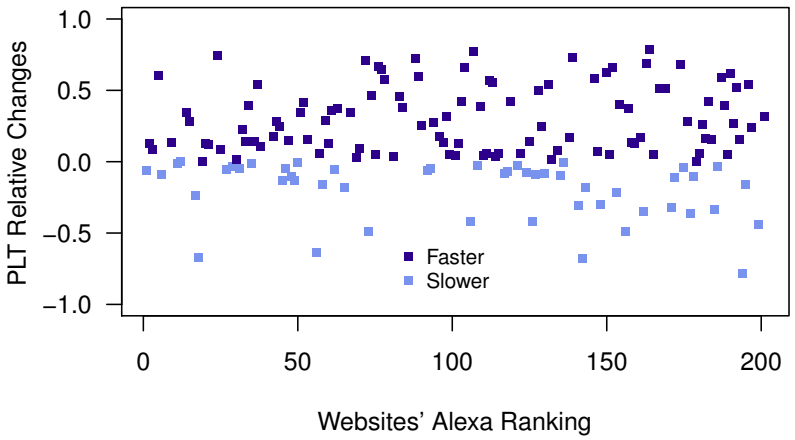


Figure 6.7: Relative changes of PLT in the presence of Ghostery against websites' ranking.

and `theladbible.com` have the highest relative increase of PLT.

We note that some websites can detect the presence of tracker-blockers and may disrupt the normal rendering of their Web pages. We use the dataset of such websites reported in [103] and identify 10 websites in our study that track the presence of tracker-blockers. However, after manually analysing these websites none are found to diverge from the normal execution. Hence, the page-load time is not influenced by the tracker-blocking detection mechanism.

6.4.1 Popular Websites

So far we have observed a variety of effects that tracker-blockers have on the performance of the websites. The arising question here is the extent to which the popular websites are affected by the tracker-blockers. It is expected that highly popular websites are managed optimally to experience minimum negative impact at the presence of tracker-blockers. To answer this question, we analyse the relative changes of the PLT when Ghostery is employed in comparison with the standard condition across websites' Alex ranking. We only consider Ghostery as it covers a larger number of trackers in comparison with AdBlock Plus.

We observe that more than half of Alexa's top-50 websites benefit from Ghostery with up to 30% reduction of the PLT (Figure 6.7). Amongst the top-50th, `yahoo.com` benefits the most with 60% faster load in comparison with the standard condition (Table 6.1). However, there are 16 websites amongst the top-50 which load slower in the presence of Ghostery of which `ebay.com` has the highest increase of PLT with 60%. We note that these are highly popular websites and as

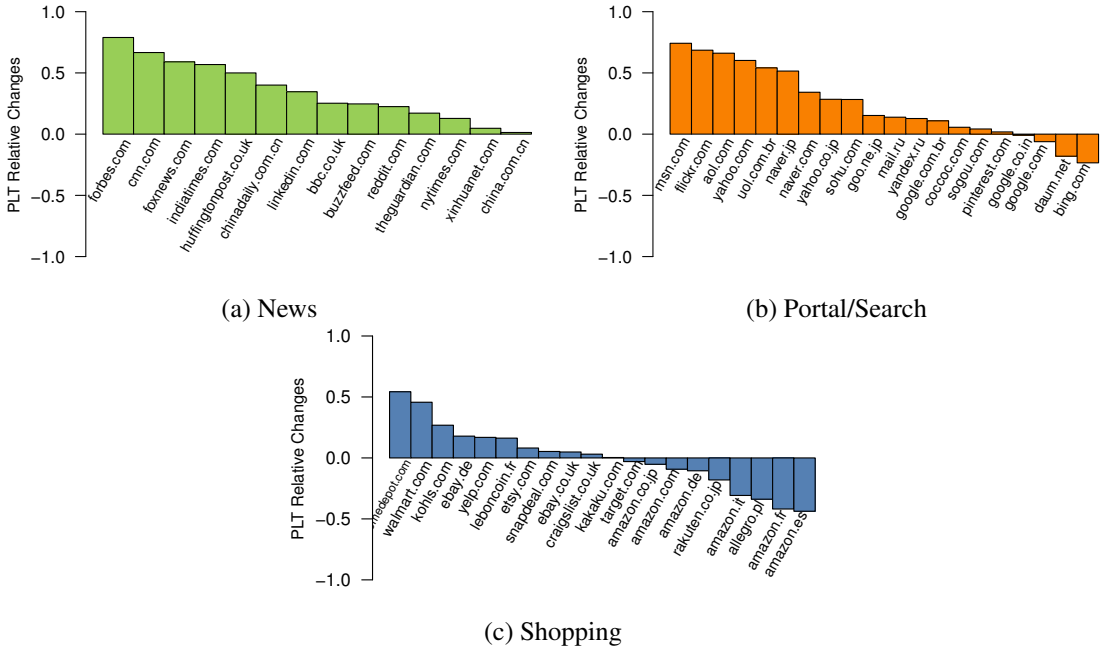
Rank	Website	Standard PLT	Ghostery PLT	Relative Change
1	google.com	2.375	2.5215	-0.06
2	facebook.com	1.988	1.7375	0.12
3	youtube.com	4.335	3.963	0.08
5	yahoo.com	10.8595	4.3245	0.60
6	amazon.com	9.601	10.488	-0.09
11	google.co.in	2.122	2.1415	-0.01
12	live.com	2.017	2.021	-0.002
14	linkedin.com	3.198	2.091	0.34
15	yahoo.co.jp	10.183	7.285	0.28
17	bing.com	0.3775	0.466	-0.23
18	ebay.com	8.6775	14.5085	-0.67
20	yandex.ru	3.249	2.8345	0.12
21	vk.com	7.496	6.612	0.11
24	msn.com	14.861	3.8325	0.74
25	instagram.com	2.7215	2.48	0.09
27	amazon.co.jp	12.205	12.8425	-0.05
30	pinterest.com	5.374	5.276	0.02
32	reddit.com	6.279	4.868	0.22
33	mail.ru	6.0575	5.219	0.14
34	paypal.com	10.9835	6.68	0.39
36	wordpress.com	3.0225	2.605	0.14

Table 6.1: Top-20 high ranking websites and the comparison of their PLT(second) under standard condition and when Ghostery is activated.

such, a higher number of users may be affected due to negative website performance. Our observation shows that as the website ranking increases (i.e., from high, to low rankings), the variance of changes on the PLT becomes larger. In other words, the influence of tracker-blockers on the performance of less popular websites is not constant. These differences can be due to different Web-development strategies taken by high ranking websites such as applying best-practices on Web performance optimisation (§6.2).

Category	No. Websites	+ vs. -
Portal/Search Engine	44	28 vs. 16
Shopping	31	15 vs. 16
News/Media	14	14 vs. 0
Entertainment	13	8 vs. 5
Computers	10	5 vs. 5
Social Networks	9	9 vs. 0
Pornography	9	4 vs. 5
Business	8	5 vs. 3
Web Advertisement	7	5 vs. 2
Streaming/Photo	7	5 vs. 2
Other	7	6 vs. 1

Table 6.2: Number of websites across different categories. The **+ vs. -** indicates the number of websites that are positively affected by Ghostery and vice versa.



(a) News (b) Portal/Search

(c) Shopping

Figure 6.8: Top-20 websites with the highest relative reduction of PLT in the presence of Ghostery.

6.4.2 Categories of Websites

We investigate how the tracker-blockers affect the performance of the websites across various categories of services. We categorise websites using the Trend Micro Site Safety Center categorisation service [88]. Our visited websites fall into 11 categories (Table 6.2) of which *Portal/Search Engine*, *Shopping* and *News/Media* appeared the most. The majority of the websites in the aforementioned categories benefit from Ghostery. Moreover, all of the websites in the *Social Networks* benefit from Ghostery. Figure 6.8 shows top-20 websites in the top-3 popular categories and their relative changes of PLT. The PLT of all of the websites in the News category reduces with `forbes.com` and `cnn.com` benefit the most with more than 50% reduction in their standard PLT (Figure 6.8a). We find 16 websites amongst the Portal/Search related websites that their PLT is affected negatively by Ghostery (6.8b). In this category different localised versions of Google search engine are of interest. We observe that the relative changes of PLT across different versions of Google are affected diversely from almost 10% reduction to 6% increase. In fact, these localised versions are not identical as they may use different resources being served from different servers and locations. The negative effect of Ghostery on the page-load time of Shopping related websites is stronger in comparison with other categories. Ghostery increases the PLT of more than half of the websites in this category (16 of 31) with more than 20%. In-

terestingly, the majority of negatively affected websites in this category are localised versions of Amazon online shopping website (Figure 6.8c). In contrast to Google, all of the versions of Amazon shopping website are negatively affected.

6.5 Related Work

There are various performance measurements focusing on the effect of network related factors and characteristics of Web pages on page-load time. Naylor *et al.* [104] investigate the impact of HTTPS on network latency. They observe a negative impact of HTTPS (above 500ms) on page-load time of about 90% of the Alexa's top-500 websites when they are loaded via mobile network, and 40% of the websites when fiber network is used. They find that TCP handshake time takes longer for half of the websites when HTTPS is used. Huang *et al.* [101] study the page-load performance under 3G network on different mobile devices. They observe that network related metrics such as round trip time and simultaneous TCP connections affect page-load performance differently across different mobile devices. Butkiewicz *et al.* [3] study the effect of different factors on page-load performance. They identify number of objects, number of JavaScript resources, Web page size and number of requests as the most correlated factors to the page-load performance. However, they observe that these factors affect various categories of websites differently. For instance, the number of images and requests have the most effect on the performance of News sites. While the total size of JavaScript objects and the number of requests are mainly important for Gaming websites. In a simulation-based study done by Ihm and Pai [2] the effect of dependencies between objects, number of simultaneous TCP connections and caching on page-load performance are studied. They simplify their simulation by ignoring network latency and browser processing time. They observe that 8 simultaneous TCP connections can reduce the page-load time by 23%. Additionally, removing dependencies between objects will reduce page-load time by 50%. These studies investigate the complex relation of page-load performance with various factors that are directly involved in the construction and delivery of a Web page. We analyse the effect of a different factor on page-load performance, namely the present tracker-blocker plug-ins that interfere with normal load of Web pages.

There are various proposals to improve page-load performance by modifying the page-load process and browser computation. Wang *et al.* [105] present a technique to re-structure the page-load

process by prioritising the parts of the Web page that needs to be loaded. Their technique is aimed to reduce the parsing-blocking latency caused by dependencies amongst JavaScript resources. This technique relies on a proxy server to pre-process a Web page and transfer a basic representation of the page's DOM to the client browser while resources with lower priority loads in the background. The Adrenaline architecture parallelizes page-loading by dividing a Web page into smaller pages and processing each page separately in parallel [106]. The WProf browser profiler [90] generates dependency graph of the browser processes. They observe that the computations needed to parse HTML tags and evaluate JavaScripts takes about 30% of page-load time.

Some recent studies focus on anti-tracker-blocking which refers to a practice in which websites detect the presence of tracker-blocker plug-ins, and coerce users to deactivate their tracker-blockers. Rafique *et al.* [107] measure anti-tracker-blocking as part of a bigger study of malicious advertisements and malware on free live-streaming services. They find 16.3% of the 1,000 examined websites use scripts that detect the presence of tracker-blockers and defeat them by forcing users to deactivate their tracker-blockers. Nithyanand *et al.* [89] identify 14 anti-tracker-blocking scripts that are employed by 6% of Alexa top-5k websites. They find that anti-tracker-blocking services were popular amongst news, blogs and entertainment websites. Interestingly, they observe that some tracker-blockers such as Ghostery and AdBlock Plus can identify anti-tracker-blocking services and counter-block them. We reveal another dimension of difficulties that users may face when trying to preserve their privacy by employing tracker-blocker plug-ins.

6.6 Conclusion

We investigated the impact of tracker-blockers on page-load performance. For this purpose, we selected two popular tracker-blocker plug-ins, Ghostery and AdBlock Plus, and measured how they affect the loading time of Alexa's top-200 websites. We observed that Ghostery and AdBlock Plus had different effects on the performance of the websites. Ghostery had a positive impact on the performance of majority of the websites (103 websites,64%) by reducing their PLT. However, AdBlock Plus improved the performance of a limited number of websites (53 websites, 33%). In line with our expectation, we observed a considerable number of websites that load slower in the presence of tracker-blockers. Ghostery increased the page-load time of 52 websites, whereas this number is almost 2 times higher in the presence of AdBlock Plus

(107 websites). Interestingly, we observed that localised versions of the same website (e.g., `google.com` vs. `googe.com.br`) are affected diversely.

Despite our initial expectation, we found highly popular websites such as `ebay.com` amongst those that were affected negatively. The page-load performance of 39% (=16) of Alexa's top-50 has reduced by 10% in the presence of Ghostery.

We note that the poor performance of websites caused by tracker-blockers can discourage users from using such tools. For instance, the extra latencies (imposed by tracker-blockers) experienced by users when using online banking facilities can make them impatient and subsequently leading users to disable their tracker-blockers. A more optimistic assumption is that users dismiss performance in favour of preserving their privacy. This assumption may be true for some users who are aware about online user tracking and concerned about their privacy. However, this assumption cannot be generalised to all users with different level of awareness regarding user tracking. Therefore, general users may have a reduced interest in using such tools. In Section §7.3, we further discuss the dilemma that users and publisher websites encounter when tracker-blockers are in use.

To conclude, the different effects of tracker-blockers highlight the complexity of today's Web in which third-parties are deeply entangled within websites. Our findings are useful for publisher websites as we recommend to take into account the user experience when tracker-blockers are used. Additionally, we revealed that tracker-blockers need to be better tuned as these plug-ins are aimed to improve user privacy with no (or minimum) negative effect on the other aspects of user browsing experience. In Section §7.3, we provide possible solutions to improve tracker-blocking plug-ins. Our work can be extended by identifying and understanding the nature of the third-party trackers that their blockage has the worst impact on the page-load performance.

Chapter 7

Conclusions and Future Directions

7.1 Summary

In this thesis, we explored the third-party tracking ecosystem from three dimensions which are 1) the geographic scope of the third-party trackers, 2) the interactions between different players and 3) privacy.

We first explored the geographic scope of the third-party trackers (Chapter §4). Previous works had reported that dominant organisations such as Google are highly present in the third-party tracking ecosystem (see Section §3.1). Considering the expanding scope and applications of third-party trackers, we expected that they would expand beyond particular countries and organisations. Our results revealed the presence of local third-party trackers in almost all the investigated countries. The international presence of third-party tracking services belonging to countries with different approaches towards user data protections and online privacy (see Section §3.2.1) adds to the complexity of this ecosystem. Moreover, the global presence of third-party trackers leads to a number of challenges for the financial and regulatory sectors. We discuss these challenges in Section §7.2.

After observing the presence of third-party trackers, we subsequently explored the possible interactions between them. In particular, we studied the nature of tracking groups sharing user-

specific IDs (Chapter §5). We identified that there exists numerous interactions between various players that share user-specific IDs. Moreover, we show that the presence of these interactions varies across different user profile conditions (logged-in vs. logged-out). Our findings highlighted the challenges faced in order to preserve users' privacy as explained in Section §7.2.

Considering the amount of user tracking that is happening on the Web, many users have turned to tracker-blocker plug-ins to mitigate the impact of third-party trackers. However, these tools may have a negative effect on user's browsing experience. We investigated the effect of tracker-blockers plug-ins on page-load performance, which is an important factor for the users (Chapter §6). We identified that blocking third-party trackers by these plug-ins can have diverse effects on the page-load performance. The diverse effects of tracker-blockers reflected on potential negative impact on users' experience when attempting to preserve their privacy.

In the remainder of this section, we provide more details about our findings on each of the aforementioned studied dimensions. Additionally, we discuss the encountered challenges through our investigation of the third-party tracking ecosystem. We conclude this chapter by providing the potential opportunities for future work.

Third-party trackers are beyond a set of global dominant players. We investigated the geographical distribution of the third-party tracking ecosystem by studying the global and regional presence of third-party trackers. The chosen websites consisted of the Alexa top-500 most popular websites of 28 countries covering five geographic regions: North America, South America, Europe, East Asia, the Middle East and Oceania. We observed the dominant presence of third-party trackers belonging to international corporations across all regions. However, we revealed the existence of local third-party trackers (third-party trackers that are physically hosted in the related region) that are even dominating global players in certain regions. For example, in North America, Europe and East Asia the presence of dominant local players was considerable. We observed a non-uniform geographic distribution of local third-party trackers across popular websites of different countries, with a dominant presence of players based in the US, Japan, Great Britain and Germany. Interestingly, local third-party trackers of some countries such as Russia seemed to be employed in popular websites of particular countries such as Turkey.

User-specific ID-sharing happens prevalently across organisations and categories of services. We studied all tracking parties that generate *user-specific IDs* and share these IDs with

other parties. We provided a first look at the nature of such tracking groups and their relation with user profiles. In our analysis, we created a dataset from both browsing histories of 129 users and active experiments. We found a significant amount of ID-sharing across different organisations providing various service categories. We even observed health related websites being part of user-specific ID-sharing groups which reflects on the possible accessibility of user sensitive information amongst the ID-sharing parties. We observed that ID-sharing happens on a large scale regardless of the user profile size and user state such as logged-in and logged-out. We believe that our analysis has revealed the huge gap between what is known about user tracking and what is done by these services.

The tracker-blockers have different effects on the page-load performance of websites. We investigated the effect of tracker-blockers on page-load performance of Alexa top-200 websites. We observed a considerable number of websites across various service categories that load slower in the presence of tracker-blockers (AdBlock Plus : 66%, Ghostery : 32%). In particular, we found shopping related websites as the most negatively affected ones. Some of the negatively affected websites took above two times longer to load in the presence of tracker-blockers. We found highly popular websites amongst those that were negatively affected which reflects on the high number of users who may experience this latency.

7.2 Challenges

Handling international third-party trackers. The presence of international third-party trackers across local websites of countries leads to certain challenges in terms of handling international flow and trade of personal information. Firstly, countries may have a different view on the nature of cyberspace. For example, Russia considers the uncontrolled flow of information as a national threat. The presence of overseas third-party trackers across popular websites of countries like Russia (e.g., we observed strong presence of German third-parties across popular Russian websites) endangers their approach towards cyber security. Another challenge here is how to retain the financial rights of different parties specially the international ones that are making money out of tracking citizens of another country. One example is the recent dispute over the inconsequential Google UK's tax (17%) whereas UK advertisers may have provided up to £5 billions of Google's sales [82]. To tackle this challenge some countries such as France have proposed a new

tax model for *personal data collection* [108]. We believe that adopting such proposals can affect the presence of the international players in this ecosystem.

Lack of transparency. The practice of user-specific ID sharing enables the involved parties to merge their user datasets on the back-end based on shared user-specific IDs, and thus extend their knowledge about user's browsing history. We cannot directly observe whether such dataset-merging happens at the back-end. In fact, not much is known about the prevalence of back-end dataset-merging which again supports the need for more transparency on the collaborations between tracking parties. A solution can be the adoption of privacy-preserving user tracking techniques that rely on trusted intermediaries for passing and merging user information.

If such collaboration between tracking parties happens (without adopting privacy-preserving techniques), a large amount of user's browsing history will be known to large number of parties involved in ID-sharing [16]. To mitigate the effect of ID-sharing, a blunt approach is to block third-party trackers. Modern browsers offer blocking third-party cookies option. Additionally, there are various tracker-blocker browser plug-ins such as Ghostery, AdBlock Plus and Privacy-Badger (see 3.4). However, these methods are not specifically targeting ID-sharing, therefore, may not necessarily block ID-sharing related traffic.

Dilemma. Tracker-blocking plug-ins have been created to answer user's frustration over plethora of invisible players collecting personal information. However, these plug-ins are perceived as a threat by publisher websites. [24] showed that the use of tracker-blockers can drop the revenue of publisher websites by 30% due to blocking advertisement entities. Our investigation revealed another possible dimension of revenue lost as a result of the reduced page-load performance. Moreover, the use of tracker-blockers comes at a price for users. Some publisher websites take discriminative actions against users of tracker-blockers by blocking them, or otherwise coercing them to disable the tracker-blockers [89]. The negative effect of tracker-blockers on page-load performance suggests another potential burden to users. The challenge is to get the right balance between the user privacy and websites performance. Therefore, the tracker-blocking plug-ins must take into account the impact of their services on the overall user experience and not irrationally block all third-party trackers. In addition, websites should apply standard optimization rules that aim to improve page-load performance (§6.2).

7.3 Future Path.

Providing measurement infrastructure. Our understanding about third-party trackers in Africa and the Middle East is still limited, perhaps, due to the lack of a stable measurement infrastructure in these regions. For example, in Africa, 8 PlanetLab nodes were deployed during our course of study of which none were available at that time. Another measurement infrastructure is RIPE Atlas [109] providing broadly deployed vantage points across the globe. However, it supports limited types of measurements. Therefore, there is a need for a geographically spread infrastructure with less limitations.

Tracker-blocking tools. While all the existing tracker-blocking plug-ins mainly rely on examining the traffic at the browser-side, we believe that adding a back-end engine performing auxiliary crawls and analysis will greatly improve these plug-ins. In fact, our approach for identifying user-specific ID-sharing is extensible as a back-end process for a new class of tracker-blocking plug-ins targeting ID-sharing. For instance to identify user-specific IDs, a series of repeated crawls of the domain which is visited by the user can be used to apply our method to distinguish user-specific IDs from non-relevant IDs. The precision of such a plug-ins in identifying user-specific IDs will gradually improve with more user engagement.

Furthermore, publisher websites have limited insight into the tracking activities of the third-party trackers embedded on their websites. The plug-ins that can identify and report the type of tracking happening on their websites will allow publishers to understand the privacy violation risks their visitors face. Hence, publishers should be accountable for the third-party tracking services they incorporate.

User experience. Considering the increasing number of tracker-blocker users, the interaction between these plug-ins and websites needs further investigation. A possible research direction is to investigate the effect of these tools on user experience. For example, different user interactions (e.g., navigating, shopping) may affect the page-load process differently and consequently the user experiences can vary. For this purpose, other metrics such as above-the-fold (shows visual completeness of a page) may be useful, although using this metric requires manual inspection unless some automatic methodology is introduced.

7.4 Final Word

The third-party trackers have received attention from different perspectives in the recent years. From a privacy point of view, our findings showed that users are being tracked extensively and unconditionally. We observed the accessibility of sensitive user information to the tracking parties which sadly has been reported in a number of previous works as well [12, 13, 9]. This unchanged situation suggests that various researches and discussions done by different communities have not had enough practical impact yet.

Our multi-dimensional analysis showed the global expansion of 2the third-party tracking ecosystem. We showed the entangled connections amongst all players, and the consequent difficulties and dilemma to protect end-users against the privacy risk 2the third-party trackers pose. Moreover, this thesis pointed to important gaps in our knowledge and understanding about various aspects of 2third-party tracking. We believe that filling these gaps will be greatly beneficial for all involved technical and legal communities, and last but not least the end-users.

Appendices

Appendix A

Automated Website Visiting

This Python software program enables us to create an instance of Firefox browser and visit websites on the browser. This software program contains: (1) the *firefoxBrowser* class managing Firefox browser functionality and (2) the *visitWebsites* function managing the process of visiting websites via the *firefoxBrowser* object.

```
import time
import webbrowser
import csv
import sys
import os
import shutil
from selenium import webdriver
from selenium.webdriver.firefox.firefox_binary import FirefoxBinary
import time
class firefoxBrowser():
  #Generates an instance of Firefox browser
  ##and installs other plug-ins such as ADBlock Plus
  def setUp(self):
    filename = "<path_to_data_collection_plug-in>"
```

```
        adbPlus_filename = "<path_to_adblock_plus>"
        browserPath = "<path_to_firefox_browser>"
        adbPlus_profile = "<profile_folder>"
        profile = webdriver.FirefoxProfile(adbPlus_profile)
        #Installs the plug-in on the Firefox
        profile.add_extension(adbPlus_filename)
        self.xvfb = Xvfb(width=1280, height=720)
        self.xvfb.start()
        self.driver = webdriver.Firefox(firefox_profile=profile,firefox_binary =
            FirefoxBinary(browserPath))

    #Opens the given URL in the browser
    def openUrl(self, thisUrl):
        driver = self.driver
        driver.get(thisUrl)

    #Closes the browser
    def closeBrowser(self):
        self.driver.close()



#Visits each website for 10 times
##Input parameter : inputFile (data type = string,
###desc. = address of the file containing a list of URLs
##Output = None
def visitWebsites(inputFile):
    i = 1 # Number of current iteration
    max = 10 # Total number of iterations
    siteStay = 60  # Duration of visiting each website
    while  i <= max:
        #Each iteration should be run from its allocated folder
```

```
    if i > 1:
      userpath = "<path_to_data_collection_plugin>"
      src = userpath + str(i) + "/persist.js"
      dst = userpath + "resources/findtracker/lib/persist.js"
      os.remove(dst)
      time.sleep(5)
      shutil.copy(src, dst)
      time.sleep(5)
      log = "no " + str(i)

    with open(inputFile,"rb") as theFile:
      line = csv.DictReader(theFile)
      count = 0
      thisBrowser = firefoxBrowser()
      thisBrowser.setUp()
      for l in line:
          url = "http://www."+l["tabDomain"]
          try:
              #Opens the given url in the browser
              thisBrowser.openUrl(url)
            time.sleep(siteStay)
            thisBrowser.closeBrowser()
            time.sleep(10) #Delay between closing and opening the browser
          except:
              print url + "," + str(sys.exc_info()[0])
              print sys.exc_info()[1]
              continue
    #Next iteration
    i = i + 1

if __name__ == "__main__":
```

```
if len(sys.argv) > 1:
 #Path to the file containing a list of URLs
 file = sys.argv[1]
 visitWebsites(file)
```

Appendix B

ADNS Method

The ADNS method described in Chapter §2 to identify third-party trackers is implemented in form of a Python software program. The software contains five functions; Above each function there is a description of the functionality, input and output.

```
import sys
import os
import commands
import csv
import tldextract

#Gets the ADNS information of a given domain name using nslookup utility tool
##and store the output in a CSV file
##Input parameter(s): domain (data type = string,
###desc. = domain name)
##Output: None
def nsLookup(domain):
    #Output of nslookup is temporarily stored in the below parameters
    origin = "" #registered origin domain name
    mail = "" #registered email address
```

```
    p = commands.getstatusoutput('nslookup -query=soa %s' % domain) #Calls
        nslookup utility tool
    try:
        p = p[1]
        arrp = p.split("\n")
        for a in arrp:
            if a.find("origin") > -1:
                origin = a.split("=")[1]
            elif a.find("mail") > -1:
                mail = a.split("=")[1]
        org2 = replaceGeneralDomains(origin,site)
        mail2 = replaceGeneralDomains(mail,site)
        outStr = site + "," + origin + "," + mail + "," + getDomainName(org2)
             + "," + getDomainName(mail2)
        wrtieIntoFile(outStr, "output/out_nsrecord","")
    except:
        wrtieIntoFile(outStr, "output/out_nsrecord","")


#Compares the obtained ADNS information with a predefined list of CDNs
##and general hosting services.
##Input parameter(s): ns (data type = string,
###desc. = The original domain name obtained from nsLookup function),
###and domain (data type = string, desc. = domain name)
##Output: String
def replaceGeneralDomains(ns,domain):
        #List of CDNs and general hosting services identified in our dataset
        general_list = ["1und1","host","cdn","dns",
                    "cloudflare","akamai","1and1",
                    "mediatemple","mailclub","moniker",
                    "hichina","dynect.net","domaincontrol.com",
                    "akam","rackspace.com","registrar-servers.com",
```

```
                    "gandi.net","name-services","softlayer.com",
                    "hyp.net","nic.ru","technorail.com",
                    "value-domain.com","loopia.se","namespace4you.de",
                    "name.com","hostmaster.netnames.net"]
    for g in general_list:
                #if it is a general domain name,
                ##use the main domain instead of ns record
                if ns.find(g) > -1:
            return domain
    return ns


#Extracts the domain name of a URL, eg., bbc.co.uk from player.bbc.co.uk
##Input parameter(s): domain (data type = string, desc. = full URL
##Output: domain name
def getDomainName(domain):
    try:
        extracted = tldextract.extract(domain)
        newDomain = "{}.{}".format(extracted.domain, extracted.suffix)
        return newDomain
    except:
        return domain


#Creates a file in CSV format
##Input parameter(s): data (data type = string),
###inputFile (data type = string, desc. = file address)
###and header (data type = string, desc. = header of the file)
##Output: None
def wrtieIntoFile(data, inputFile, header = ""):
    #if you don't want header leave this parameter empty
    if header:
        if not os.path.isfile(inputFile):
```

```
            with open(inputFile, 'a') as the_file:
                the_file.write(header)
                the_file.write(os.linesep)
    with open(inputFile, 'a') as the_file:
        the_file.write(data.encode('utf-8').strip())
        the_file.write(os.linesep)


#Gets the ADNS information of all URLs in the inputFile
##Input parameter(s): inputFile (data type = string, desc. = file address;
###the input file MUST have a header row calling "url"),
##Output: None
def getAdnsInfo(inputFile):
    with open(inputFile, "rb") as theFile:
        line = csv.DictReader(theFile)
        for l in line:
            nsLookup(l["url"])
```

Bibliography

[1] D. Crane, E. Pascarello, and D. James. *Ajax in Action*. Manning Publications Co., Greenwich, CT, USA, 2005.

[2] Sunghwan Ihm and Vivek S. Pai. Towards understanding modern web traffic. In *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference*, IMC '11, pages 295–312, New York, NY, USA, 2011. ACM.

[3] Michael Butkiewicz, Harsha V. Madhyastha, and Vyas Sekar. Understanding website complexity: Measurements, metrics, and implications. In *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference*, IMC '11, pages 313–328, New York, NY, USA, 2011. ACM.

[4] Nsa using google's online ad tracking tools to spy on web users. `http://threatpost.com/nsa-using-google-non-advertising-cookie-to-spy/`.

[5] Apostolis Zarras, Alexandros Kapravelos, Gianluca Stringhini, Thorsten Holz, Christopher Kruegel, and Giovanni Vigna. The dark alleys of madison avenue: Understanding malicious advertisements. In *Proceedings of the 2014 Conference on Internet Measurement Conference*, IMC '14, pages 373–380, New York, NY, USA, 2014. ACM.

[6] Balachander Krishnamurthy and Craig Wills. Privacy diffusion on the web: A longitudinal perspective. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 541–550, New York, NY, USA, 2009. ACM.

[7] Adam Lerner, Anna Kornfeld Simpson, Tadayoshi Kohno, and Franziska Roesner. Internet jones and the raiders of the lost trackers: An archaeological study of web tracking from 1996 to 2016. In *25th USENIX Security Symposium (USENIX Security 16)*, Austin, TX, August 2016. USENIX Association.

[8] Claude Castelluccia, Stéphane Grumbach, and Lukasz Olejnik. Data Harvesting 2.0: from the Visible to the Invisible Web. In *The Twelfth Workshop on the Economics of Information Security*, Washington, DC, United States, June 2013. Allan Friedman.

[9] Balachander Krishnamurthy, Konstantin Naryshkin, and Craig Wills. Privacy leakage vs. protection measures: the growing disconnect. In *Proceedings of the Web*, volume 2, pages 1–10, 2011.

[10] Jonathan Mayer. Tracking the trackers: Where everybody knows your username. `http://cyberlaw.stanford.edu/blog/2011/10/tracking-trackers-where-everybody-knows-your-username/`.

[11] Abdelberi Chaabane, Yuan Ding, Ratan Dey, Mohamed Ali Kaafar, and Keith Ross. A Closer Look at Third-Party OSN Applications: Are They Leaking Your Personal Information? In *Passive and Active Measurement conference (2014)*, Los Angeles, États-Unis, March 2014. Springer.

[12] Craig E. Wills and Can Tatar. Understanding what they do with what they know. In *Proceedings of the 2012 ACM Workshop on Privacy in the Electronic Society*, WPES '12, pages 13–18, New York, NY, USA, 2012. ACM.

[13] Delfina Malandrino, Andrea Petta, Vittorio Scarano, Luigi Serra, Raffaele Spinelli, and Balachander Krishnamurthy. Privacy awareness about information leakage: Who knows what about me? In *Proceedings of the 12th ACM workshop on Workshop on privacy in the electronic society*, pages 279–284. ACM, 2013.

[14] Balachander Krishnamurthy and Craig E. Wills. Privacy leakage in mobile online social networks. In *Proceedings of the 3rd Wonference on Online Social Networks*, WOSN'10, pages 4–4, Berkeley, CA, USA, 2010. USENIX Association.

[15] Arvind Narayanan and Vitaly Shmatikov. De-anonymizing social networks. In *Proceedings of the 2009 30th IEEE Symposium on Security and Privacy*, SP '09, pages 173–187, Washington, DC, USA, 2009. IEEE Computer Society.

[16] Gunes Acar, Christian Eubank, Steven Englehardt, Marc Juarez, Arvind Narayanan, and

Claudia Diaz. The web never forgets: Persistent tracking mechanisms in the wild. pages 674–689, 2014.

[17] Arpita Ghosh and Aaron Roth. Selling privacy at auction. In *Proceedings of the 12th ACM Conference on Electronic Commerce*, EC '11, pages 199–208, New York, NY, USA, 2011. ACM.

[18] F Stuart Chapin III, Pamela A Matson, and Peter Vitousek. *Principles of terrestrial ecosystem ecology*. Springer Science & Business Media, 2011.

[19] Ryan McCormack. Digital ecosystems: A framework for online business. `http://bitstrategist.com/2011/06/digital-ecosystems-a-framework-for-online-business/`.

[20] Robert H. Sloan and Richard Warner. *Unauthorized Access: The Crisis in Online Privacy and Security*. CRC Press, Inc., Boca Raton, FL, USA, 1st edition, 2013.

[21] Thomas Roessler. Tracking controversy. `https://www.w3.org/blog/2012/06/tracking-controversy/`.

[22] Nigel Goldenfeld and Leo P Kadanoff. Simple lessons from complexity. *science*, 284(5411):87–89, 1999.

[23] M. Falahrastegar, H. Haddadi, S. Uhlig, and R. Mortier. The rise of panopticons: Examining region-specific third-party web tracking. In *Traffic Monitoring and Analysis*. Springer Berlin Heidelberg, 2014.

[24] Phillipa Gill, Vijay Erramilli, Augustin Chaintreau, Balachander Krishnamurthy, Konstantina Papagiannaki, and Pablo Rodriguez. Follow the money: Understanding economics of online aggregation and advertising. In *Proceedings of the 2013 Conference on Internet Measurement Conference*, IMC '13, pages 141–148, New York, NY, USA, 2013. ACM.

[25] Hu Fu, Patrick Jordan, Mohammad Mahdian, Uri Nadav, Inbal Talgam-Cohen, and Sergei Vassilvitskii. Ad auctions with data. In *Proceedings of the 5th International Conference on Algorithmic Game Theory*, SAGT'12, pages 168–179, Berlin, Heidelberg, 2012. Springer-Verlag.

[26] De Filippi Primavera. Taxing the cloud.
https://policyreview.info/articles/analysis/taxing-cloud-introducing-new-taxation-system-data-collection.

[27] Clicking for gold.
http://www.economist.com/news/business/21594995-taxmen-are-doing-whatever-they-can-squeeze-more-online-businesses-patch-up-job/.

[28] Toby Mendel, Andrew Puddephatt, Ben Wagner, Dixie Hawtin, and Natalia Torres. *Global survey on internet privacy and freedom of expression*. UNESCO, 2012.

[29] William J Long and Marc Pang Quek. Personal data privacy protection in an age of globalization: the us-eu safe harbor compromise. *Journal of European Public Policy*, 9(3):325–344, 2002.

[30] Dereje Yimam and Eduardo B. Fernandez. A survey of compliance issues in cloud computing. *Journal of Internet Services and Applications*, 7(1):5, 2016.

[31] David S. Evans. The online advertising industry: Economics, evolution, and privacy. *Journal of Economic Perspectives*, 23(3):37–60, 2009.

[32] Conversion tracking guide.
https://support.google.com/adxbuyer/answer/165288?hl=en.

[33] MeMe Jacobs Rasmussen. Adobe position paper on privacy and tracking. In *W3C Workshop on Web Tracking and User Privacy*, March 2011.

[34] Introduction to ga.js (legacy). https://developers.google.com/analytics/devguides/collection/gajs/#disable.

[35] Google analytics opt-out browser add-on. https://tools.google.com/dlpage/gaoptout.

[36] Facebook's audience network.
https://developers.facebook.com/products/app-monetization/audience-network/.

[37] Franziska Roesner, Tadayoshi Kohno, and David Wetherall. Detecting and defending against third-party tracking on the web. In *Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation*, NSDI'12, pages 12–12, Berkeley, CA, USA, 2012. USENIX Association.

[38] Gunes Acar, Christian Eubank, Steven Englehardt, Marc Juarez, Arvind Narayanan, and Claudia Diaz. The web never forgets: Persistent tracking mechanisms in the wild. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, CCS '14, pages 674–689, New York, NY, USA, 2014. ACM.

[39] Ashkan Soltani, Shannon Canty, Quentin Mayo, Lauren Thomas, and Chris Jay Hoofnagle. Flash cookies and privacy. In *AAAI Spring Symposium: Intelligent Information Privacy Management*, 2010.

[40] Aleecia M McDonald and Lorrie Faith Cranor. Survey of the use of adobe flash local shared objects to respawn http cookies, a. *ISJLP*, 7:639, 2011.

[41] Peter Eckersley. How unique is your web browser? In *Proceedings of the 10th International Conference on Privacy Enhancing Technologies*, PETS'10, pages 1–18, Berlin, Heidelberg, 2010. Springer-Verlag.

[42] Lukasz Olejnik, Claude Castelluccia, and Artur Janc. Why Johnny Can't Browse in Peace: On the Uniqueness of Web Browsing History Patterns. In *5th Workshop on Hot Topics in Privacy Enhancing Technologies (HotPETs 2012)*, Vigo, Espagne, July 2012.

[43] Jonathan Mayer. Tracking the trackers: Self-help tools. `https://cyberlaw.stanford.edu/blog/2011/09/tracking-trackers-self-help-tools/`.

[44] Robert J. Walls, Shane S. Clark, and Brian Neil Levine. Functional privacy or why cookies are better with milk. In *Proceedings of the 7th USENIX Conference on Hot Topics in Security*, HotSec'12, pages 11–11, Berkeley, CA, USA, 2012. USENIX Association.

[45] Privacybadger. `https://www.eff.org/privacybadger`.

[46] Saranga Komanduri, Richard Shay, Greg Norcie, and Blase Ur. Adchoices-compliance with online behavioral advertising notice and choice requirements. *ISJLP*, 7:603, 2011.

[47] Nai consumer opt-out. http://www.networkadvertising.org/choices/.

[48] Do not track. http://donottrack.us/.

[49] Gaurav Aggarwal, Elie Bursztein, Collin Jackson, and Dan Boneh. An analysis of private browsing modes in modern browsers. In *Proceedings of the 19th USENIX Conference on Security*, USENIX Security'10, pages 6–6, Berkeley, CA, USA, 2010. USENIX Association.

[50] Saikat Guha, Bin Cheng, and Paul Francis. Privad: Practical privacy in online advertising. In *Proceedings of the 8th USENIX Conference on Networked Systems Design and Implementation*, NSDI'11, pages 13–13, Berkeley, CA, USA, 2011. USENIX Association.

[51] Vincent Toubiana, Arvind Narayanan, Dan Boneh, Helen Nissenbaum, and Solon Barocas. Adnostic: Privacy preserving targeted advertising. In *NDSS*, 2010.

[52] Matthew Fredrikson and Ben Livshits. Repriv: Re-envisioning in-browser privacy. Technical report, May 2011.

[53] Mikhail Bilenko and Matthew Richardson. Predictive client-side profiles for personalized advertising. In *Proceedings of the 17th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD-2011)*, San Diego, CA, USA, August 2011.

[54] Limin Wang, Kyoung Soo Park, Ruoming Pang, Vivek Pai, and Larry Peterson. Reliability and security in the codeen content distribution network. In *Proceedings of the Annual Conference on USENIX Annual Technical Conference*, ATEC '04, pages 14–14, Berkeley, CA, USA, 2004. USENIX Association.

[55] Omb memorandum 07-16 safeguarding against and responding to the breach of personally identifiable information. https://www.whitehouse.gov/sites/default/files/omb/memoranda/fy2007/m07-16.pdf.

[56] Erika McCallister, Timothy Grance, and Karen A. Scarfone. Sp 800-122. guide to protecting the confidentiality of personally identifiable information (pii). Technical report, Gaithersburg, MD, United States, 2010.

[57] Constance Gustke. Which countries are better at protecting privacy?

http://www.bbc.com/capital/
story/20130625-your-private-data-is-showing.

[58] Data protection laws of the world - United States.
https://www.dlapiperdataprotection.com/
index.html?t=online-privacy&c=US.

[59] Eu directive 95/46/ec - the data protection directive.
https://www.dataprotection.ie/docs/
EU-Directive-95-46-EC-Chapter-1/92.htm#2.

[60] Australian privacy principle 5 — notification of the collection of personal information.
http://www.oaic.gov.au/images/documents/privacy/
engaging-with-you/current-privacy-consultations/
Draft-APP-Guidelines-2013/Draft_APP_Guidelines_Chapter_5_
_APP_5.pdf.

[61] Data protection laws of the world - Argentina.
https://www.dlapiperdataprotection.com/
index.html?t=enforcement&c=AR.

[62] An overview of Turkey's new data protection law.
http://privacylaw.proskauer.com/2016/04/articles/
international/an-overview-of-turkeys-new-data-protection-law/.

[63] Data protection laws of the world - South Korea. https://www.
dlapiperdataprotection.com/index.html?t=definitions&c=KR.

[64] Data protection laws of the world - China.
https://www.dlapiperdataprotection.com/
index.html?t=enforcement&c=CN.

[65] Emily Steel. A web pioneer profiles users by name. http://online.wsj.com/
news/articles/SB10001424052702304410504575560243259416072/.

[66] Arnold Roosendaal. Facebook tracks and traces everyone: Like this! *Tilburg Law School Legal Studies Research Paper Series*, 2011.

[67] Company bypasses cookie-deleting consumers. http://www.informationweek.com/company-bypasses-cookie-deleting-consumers/d/d-id/1031518?

[68] Nick Cubrilovic. Persistent and unblockable cookies using http headers. https://www.nikcub.com/posts/persistent-and-unblockable-cookies-using-http-headers-2/.

[69] Mika Ayenson, Dietrich J. Wambach, Ashkan Soltani, Nathan Good, and Chris J. Hoofnagle. Flash Cookies and Privacy II: Now with HTML5 and ETag Respawning. *Social Science Research Network Working Paper Series*, July 2011.

[70] Samy Kamkar. Evercookie never forget. http://samy.pl/evercookie/.

[71] Keaton Mowery and Hovav Shacham. Pixel perfect: Fingerprinting canvas in HTML5. In Matt Fredrikson, editor, *Proceedings of W2SP 2012*. IEEE Computer Society, May 2012.

[72] Nick Nikiforakis, Alexandros Kapravelos, Wouter Joosen, Christopher Kruegel, Frank Piessens, and Giovanni Vigna. Cookieless monster: Exploring the ecosystem of web-based device fingerprinting. In *Proceedings of the 2013 IEEE Symposium on Security and Privacy*, SP '13, pages 541–555, Washington, DC, USA, 2013. IEEE Computer Society.

[73] Ghostery. https://www.ghostery.com/.

[74] Pedro Leon, Blase Ur, Richard Shay, Yang Wang, Rebecca Balebako, and Lorrie Cranor. Why johnny can't opt out: A usability evaluation of tools to limit online behavioral advertising. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 589–598, New York, NY, USA, 2012. ACM.

[75] The 2015 ad blocking report. https://pagefair.com/blog/2015/ad-blocking-report/.

[76] Christopher Soghoian. Targeted advertising cookie opt-out (taco). http://paranoia.dubfire.net/2009/07/taco-20-released.html/.

[77] The internet engineering task force. https://www.ietf.org/.

[78] Internet Engineering Task Force (IETF). Rfc6973: Privacy considerations for internet protocols.
https://tools.ietf.org/html/rfc6973.

[79] Internet Engineering Task Force (IETF). Rfc7258: Pervasive monitoring is an attack.
https://tools.ietf.org/html/rfc7258.

[80] Marjan Falahrastegar, Hamed Haddadi, Steve Uhlig, and Richard Mortier. The rise of panopticons: Examining region-specific third-party web tracking. In Alberto Dainotti, Anirban Mahanti, and Steve Uhlig, editors, *Traffic Monitoring and Analysis*, volume 8406 of *Lecture Notes in Computer Science*, pages 104–114. Springer Berlin Heidelberg, 2014.

[81] Marjan Falahrastegar, Hamed Haddadi, Steve Uhlig, and Richard Mortier. *Tracking personal identifiers across the web*, volume 9631 of *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pages 30–41. Springer Verlag, 2016.

[82] Can Google say it payes close to the rate of UK corporation tax?
https://www.theguardian.com/technology/2016/feb/11/can-google-say-it-pays-close-to-the-rate-of-uk-corporation-tax.

[83] Firefox tracker.
http://www.eecs.qmul.ac.uk/~marjan/experiment/firefox_tracker.zip.

[84] Observer notifications.
https://developer.mozilla.org/en/docs/Observer_Notifications.

[85] Maxmind. http://dev.maxmind.com/.

[86] Abine. https://www.abine.com.

[87] Collusion Firefox add on. Collusion firefox add-on. http://collusion.toolness.org/.

[88] Trend micro. http://global.sitesafety.trendmicro.com.

[89] Rishab Nithyanand, Sheharbano Khattak, Mobin Javed, Narseo Vallina-Rodriguez, Marjan Falahrastegar, Julia E. Powles, Emiliano De Cristofaro, Hamed Haddadi, and Steven J. Murdoch. Adblocking and counter blocking: A slice of the arms race. In *6th USENIX Workshop on Free and Open Communications on the Internet (FOCI 16)*. USENIX Association, August 2016.

[90] Xiao Sophia Wang, Aruna Balasubramanian, Arvind Krishnamurthy, and David Wetherall. Demystifying page load performance with wprof. In *Presented as part of the 10th USENIX Symposium on Networked Systems Design and Implementation (NSDI 13)*, pages 473–485. USENIX, 2013.

[91] Is a slow website costing you sales? `http://www.rickwhittington.com/blog/is-a-slow-website-costing-you-sales/`.

[92] 17 statistics to sell web performance optimization. `http://www.guypo.com/17-statistics-to-sell-web-performance-optimization/`.

[93] Xiao Sophia Wang, Arvind Krishnamurthy, and David Wetherall. Speeding up web page loads with shandian. In *13th USENIX Symposium on Networked Systems Design and Implementation (NSDI 16)*, pages 109–122. USENIX Association, 2016.

[94] *Silo: Exploiting JavaScript and DOM Storage for Faster Page Loads*, WebApps'10, Berkeley, CA, USA, 2010. USENIX Association.

[95] Pagespeed insights rules. `https://developers.google.com/speed/docs/insights/rules`.

[96] Adblock plus. `https://adblockplus.org/`.

[97] The best free and paid pop-up and ad blocker for browsing the internet. `https://www.pcwdld.com/pop-up-and-ad-blocker`.

[98] The top 5 most effective ad blockers in 2016 - glitch. `http://www.glitch.news/2016-02-01-the-top-5-most-effective-ad-blockers-in-2016.html`.

[99] W3c navigation timing. `https://www.w3.org/TR/navigation-timing/`.

[100] Top retail websites not getting faster: Average web page load time is 7.25 seconds. http://marketingland.com/retail-website-load-times-continue-to-decline-with-a-22-decrease-during-the-last-year-37604.

[101] Junxian Huang, Qiang Xu, Birjodh Tiwana, Z. Morley Mao, Ming Zhang, and Paramvir Bahl. Anatomizing application performance differences on smartphones. In *Proceedings of the 8th International Conference on Mobile Systems, Applications, and Services*, MobiSys '10, pages 165–178, New York, NY, USA, 2010. ACM.

[102] The best UK home broadband ISPs for 2016. http://www.ispreview.co.uk/index.php/2016/01/the-uk-best-broadband-isps-for-2016-ispreview-editor-picks.html.

[103] Rishab Nithyanand, Sheharbano Khattak, Mobin Javed, Narseo Vallina-Rodriguez, Marjan Falahrastegar, Julia E. Powles, Emiliano De Cristofaro, Hamed Haddadi, and Steven J. Murdoch. Ad-blocking and counter blocking: A slice of the arms race. *USENIX Free and Open Communications on the Internet (USENIX FOCI)*, 2016.

[104] David Naylor, Alessandro Finamore, Ilias Leontiadis, Yan Grunenberger, Marco Mellia, Maurizio Munafò, Konstantina Papagiannaki, and Peter Steenkiste. The cost of the "s" in https. In *Proceedings of the 10th ACM International on Conference on Emerging Networking Experiments and Technologies*, CoNEXT '14, pages 133–140, New York, NY, USA, 2014. ACM.

[105] Xiao Sophia Wang, Arvind Krishnamurthy, and David Wetherall. Speeding up web page loads with shandian. In *Proceedings of the 13th Usenix Conference on Networked Systems Design and Implementation*, NSDI'16, pages 109–122, Berkeley, CA, USA, 2016. USENIX Association.

[106] Haohui Mai, Shuo Tang, Samuel T. King, Calin Cascaval, and Pablo Montesinos. A case for parallelizing web pages. In *Proceedings of the 4th USENIX Conference on Hot Topics in Parallelism*, HotPar'12, Berkeley, CA, USA, 2012. USENIX Association.

[107] M. Zubair Rafique, Tom van Goethem, Wouter Joosen, Christophe Huygens, and Nick Nikiforakis. It's free for a reason: Exploring the ecosystem of free live streaming services.

In *23nd Annual Network and Distributed System Security Symposium, NDSS 2016, San Diego, California, USA, February 21-24, 2016*, 2016.

[108] France wants to tax facebook, google personal data collection. http://marketingland.com/france-wants-to-tax-facebook-google-personal-data-collection-31196.

[109] Network monitor. https://atlas.ripe.net/about/measurements/.